Lower Bounds on the LTE-A Average Random Access Delay Under Massive M2M Arrivals

Mehmet Koseoglu, *Member, IEEE*

***Abstract*— Rapid growth of machine-to-machine (M2M) communications necessitates the reevaluation of the Long Term Evolution-Advanced (LTE-A) performance, since the current standard is not optimized for intensive M2M traffic. A serious issue is that massive M2M arrivals can overload the LTE-A random access channel, resulting in a significant access delay. There have been a number of proposals to control this overload; however, there are no studies on the mathematical characterization of delay bounds to the best of our knowledge. Here, we derive lower bounds for the LTE-A average random access delay for both a regular traffic pattern (uniformly distributed arrivals) and for a traffic pattern, indicating a serious congestion (beta-distributed arrivals). The proposed delay bounds, which predict the minimum delay with less than 6% error, present the fundamental limits of delay that can be achieved by a practical load-balancing algorithm. This paper is also one of the first attempts toward the mathematical analysis of beta-distributed arrivals. We also analyze the effect of estimation accuracy, frequency of random access opportunities, and the number of preambles on the access delay. We show that it is possible to reduce the access delay by several orders of magnitude using an appropriate configuration of these system parameters.**

***Index Terms*— Machine-to-machine communications, 4G mobile communication, Radio access networks, Multiaccess communication.**

I. Introduction

ALONG with the growing adoption of the Internet of Things vision, the number of smart devices is estimated to reach 50 billion at the end of the decade [1] and machine-to-machine communications (M2M) is expected to grow rapidly. M2M traffic differs significantly from human-generated traffic: Most M2M devices infrequently upload small-sized sensing information as opposed to humans who mostly download web content [2]. Although M2M devices need to transmit small amounts of data, communication infrastructure may get congested if a huge number of M2M devices attempt to access the network near-simultaneously [3]. Hence the performance of existing cellular standards has to be evaluated for this emerging type of traffic.

The main impact of M2M communications on the 3GPP architecture is on the radio access network [4]. The major issue is the possible traffic surges of M2M traffic due to massive numbers of M2M devices deployed over an area. Synchronized behavior of these devices due to some external event, e.g. a power recovery, periodic measurement reporting, disaster alarms, can severely congest the radio access network preventing the channel access for most M2M devices.

The congestion occurs during the random access phase where devices request channel access from the base station by transmitting a randomly selected preamble over a shared channel. Since there is a limited number of preambles and the transmissions are sent over a shared channel, more than one node may select the same preamble which is called as a collision. In this case, the colliding nodes cannot be granted channel access and need to retry random access. If a large number of devices try to access the channel in a short time, preamble collisions increase significantly resulting in very large access delays. Besides, the repeated transmission attempts increase the energy consumption of M2M devices most of which will be energy-constrained.

To prevent such a congestion, there have been numerous proposals for the load control of the LTE random access channel [5]–[10]. Most proposals are based on broadcasting an access probability to the network to prevent nodes from accessing the channel according to some probability. Determination of the optimum access probability has been the main focus of these studies as the base station does not have perfect information about the number of nodes accessing the channel. Due to this lack of information, all load control algorithms are inevitably sub-optimal.

In this paper, we derive closed-form lower bounds for LTE-A average random access delay under massive M2M arrivals. These bounds determine the minimum delay that can be achieved by a practical load control algorithm. We derive these bounds for both uniform M2M arrivals and for highly synchronized (beta-distributed) M2M arrivals during a congestion. Using these bounds, we investigate the effect of the system parameters and the attributes of the arrival distribution on the access delay. The contributions of the paper can be summarized as follows:

- We propose closed-form expressions of minimum delay which predicts the average delay with less than 6% error for both uniformly-distributed and beta-distributed massive arrivals. Beta distribution has been suggested to model the M2M overload traffic distribution by 3GPP [6] but it is known to be mathematically intractable [11]. We analyze the delay performance of beta-distributed traffic by applying a triangular approximation; and, to

Manuscript received December 2, 2015; revised February 6, 2016 and March 18, 2016; accepted April 1, 2016. Date of publication April 5, 2016; date of current version May 13, 2016. This work is supported by the Science and Research Council of Turkey (Tubitak) under Project EEEAG-115E459 and by Hacettepe University Scientific Research Coordination Unit under project FDS-2015-7507. The associate editor coordinating the review of this paper and approving it for publication was L. Badia.

The author is with the Department of Computer Engineering, Hacettepe University, Ankara 06800, Turkey (e-mail: mkoseoglu@cs.hacettepe.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2016.2550526

the best of our knowledge, this is one of the first attempts towards its algebraic analysis. The proposed bounds for the uniformly-distributed traffic are valid for any given number of nodes; whereas, the bounds derived for the beta-distributed arrivals are valid either when the channel is stable or when it is seriously congested.

- We show that beta-distributed arrivals may experience up to 300 times more average delay in comparison to the uniformly distributed arrival of an equal number of nodes. This result implies that the network should be significantly over-provisioned due to the possible synchronization among nodes.
- We show that the accuracy in the estimation of the number of nodes in the network is not crucial as long as the network does not operate close to its capacity. However, if the network is operating close its capacity limits, an estimation error may cause the network to become unstable and may result in excessive delays.
- We show that it is possible to obtain several orders of reduction in delay by increasing the frequency of random access opportunities or the number of preambles devoted for M2M communications if such an increase in capacity results in the stabilization of the random access channel.
- We also consider the effect of backoff parameters and show that the average delay approaches to the proposed bounds as the maximum allowable number of backoffs increases. Hence, the proposed bounds can also be used as a benchmark for backoff parameter optimization.

We derive the lower bounds by analyzing the delay performance of an optimum load controller which exactly knows the number of backlogged nodes in the network and reacts to this load by broadcasting an access probability to all nodes. The delay performance of the optimum load controller presents a lower bound for practical load control algorithms which have imperfect information about the channel backlog. Proposed bounds can be used in optimizing system parameters, e.g. to determine the optimum frequency of random access opportunities or the number of preambles reserved for M2M communications depending on the expected arrival pattern of M2M devices.

We proposed bounds both for uniformly distributed arrivals indicating a regular traffic pattern and for beta-distributed arrivals which indicates a serious congestion in the network. Beta distribution has been used to model M2M arrivals during a serious congestion by 3GPP but its algebraic analysis is known to be difficult. We proposed bounds for the Beta-distributed traffic (which are valid either when the channel is stable or seriously loaded) by suggesting a triangular approximation to the Beta distribution. Success of this approximation suggests that the triangular distribution can be used in the analysis of the random access channel instead of the Beta distribution.

The rest of the paper is organized as follows: We explain the related work in the next section. Then, we describe our system model in Sec. III and we define the optimum load controller in Sec. IV and analyze its delay performance in Sec. V. We analyze the effect of estimation error and capture effect on the channel capacity in Secs. VI and VII, respectively. We compare the proposed analysis against simulation results in Sec. VIII and present our conclusions in Sec. IX.

II. Related Work

Providing wireless access to a massive number of smart devices is a challenging problem. Although there have been several alternative proposals for the M2M network architecture, the most promising candidate is the existing cellular network system due to its large coverage [11]. Hence, there is a continuing effort in the 3GPP for defining M2M service requirements and application scenarios [12], [13]. M2M traffic shows different characteristics than H2H traffic: Most M2M devices infrequently transmit a small message such as measurement reports by a smart meter [2]. Since M2M devices have a low traffic demand, there will be a larger number of M2M devices per base station in comparison to H2H devices.

Dense deployment of M2M devices may cause occasional congestions in the radio access network [4]. The channel access phase starts with a random access procedure in which the node transmits a randomly selected preamble over a shared channel. If too many nodes request access through the random access channel, a preamble may be selected by more than one node. In that case, the nodes cannot be granted channel access by the base station. Such collisions necessitate the retransmission of the preamble which causes significant access delays. There has been a growing interest in modeling the collision probability and delay of the random access channel [14]–[22].

There have also been numerous proposals for load control of the LTE random access channel. There are two legacy schemes used by 3GPP for load control namely, access class barring (ACB) and extended access barring (EAB): In ACB, the base station broadcasts an access probability p and each node selects a random number and transmits a preamble if the selected number is larger than p, otherwise the node waits for a random amount [4], [23], [24]. In EAB, the nodes are barred from channel access depending on their access class providing some sort of service differentiation [25].

Previous studies focused mostly on the optimization of ACB access probability using estimation and control techniques. Adaptive determination of the barring probability in ACB has been considered in [5] using a PID controller. Stabilization of the random access channel using channel history has been studied by Wu et al. [6]. Galinina et al. proposed a stable control procedure for the random access channel [7]. A heuristic algorithm to update access probability is proposed by Duan et al. [8]. Arouk et al. proposed a filtering based adaptation of the access probability [9]. A Kalman filtering approach has been investigated in [10].

Although there have been numerous proposals for load control and delay analysis, the fundamental limits of delay have not been previously investigated to the best of our knowledge. Here we propose closed-form expressions for the minimum delay that can be achieved by a practical load controller. Proposed expressions allow analytical investigation of the effect of various system parameters and they provide a benchmark for practical load control algorithms.

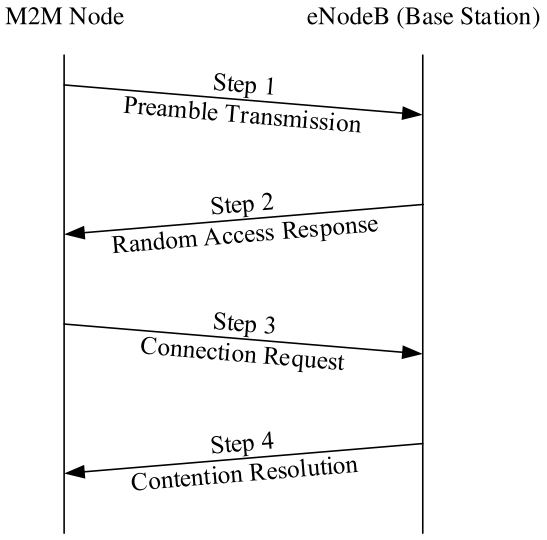

Fig. 1. LTE contention-based random access procedure.

Aside from these studies, there are also various studies aiming to provide service differentiation and to adapt the LTE access procedure for M2M communications. A class-dependent load control which requires classifying machines according to their delay-tolerance and serving them according to their priority is proposed by Cheng et al. [26]. Another load balancing method based on instructing nodes their next channel slots after a successful connection has been proposed in [27]. Reserving groups of preambles for H2H and M2M traffic has also been proposed [28]. Adaptive sharing between preambles between non-contention based and contention based transmissions is proposed in [29]. There is also a proposal to increase the number of preambles by using codewords [30].

For more details, readers may refer to several surveys on the topic [2], [3], [31].

III. System Model

A. LTE Random Access Procedure

In the LTE standard, the spectrum is divided into 10 ms frames and frames are further divided into 1 ms subframes. Several of these subframes are reserved for the random access procedure which we call as random access opportunities (RAO). The RAOs can be presented in varying frequencies ranging from once in 1 ms to once in 20 ms depending on the PRACH configuration.

The contention-based random access procedure of the LTE consists of four steps [32] as shown in Fig. 1. In the first step, the node randomly selects one of the 54 preambles reserved for contention-based random access and transmits the selected preamble at a RAO. In the second step, after the detection of the preamble, a random access response (RAR) is transmitted by the base station (eNodeB) consisting of a timing correction, scheduling grant and a temporary identity for the successfully received preambles. Using this information, the node transmits its identity at the third step. In the fourth step, the eNodeB transmits a contention resolution message.

If only a single node transmits a preamble at Step 1, the preamble will be detected with a high probability by the eNodeB. Then, the remaining steps should be completed without failure (unless a loss due to PDCCH congestion or a physical layer impairment occurs). If more than one node selects the same preamble and transmits at the same RAO, there are two possibilities: a) The transmissions will interfere with each other and the collision will be detected by the base station. In this case, the base station will not transmit RAR in Step 2. b) The transmissions by multiple nodes will overlap and the collision is not detected by the eNodeB. In this case, all nodes transmitting the same preamble will receive the same RAR at Step 2 and a collision will occur at Step 3 when the nodes transmit their identity at the same scheduled interval. In this case, if the eNodeB detects one of the transmissions, it will send a contention resolution message at Step 4 to indicate which of the nodes was granted channel access.

Most studies in the literature assume that the base station detects the collision at Step 1 [5]–[10]. There are also several studies which deals with the second case where the collision occurs at Step 3 [20]. There is a recent proposal for random access to ensure that the collisions are detected at Step 1 [33]. As most of the studies in the literature, we assume that the preamble collisions are detected during the first step.

B. System Model

We consider the arrival of N nodes to the LTE random access channel with an arrival rate of $\lambda(t)$ over a bounded time interval $[0, T]$. The base station presents the RAOs with an average time separation of τ and it broadcasts an attempt probability q before each RAO. At a RAO, each backlogged node selects a random number between 0 and 1 and transmits a preamble only if the selected number is less than q. The transmitted preamble is selected randomly from the set of p preambles reserved for M2M traffic. If two or more nodes select the same preamble, all of those transmissions collide at the receiver and cannot be correctly decoded by the base station. In the case of a collision, all colliding nodes remain backlogged. We assume that the newly arriving nodes join the pool of backlogged nodes as in [7] and the backlogged nodes repeat the procedure to transmit a preamble in the next RAO using the new q value announced by the base station.

We define the access delay of a node as the time difference between the first RAO after its arrival and the RAO at which it successfully departs. By this definition, the access delay of a node which departs at the first RAO after its arrival is zero.

Our analysis deals with a simplified model of the random access procedure: We assume that a node remains backlogged until it successfully accesses the channel; that is, we omit the upper bound on preamble transmission attempts and assume that the backoff interval is zero similar to some other analyses [7], [17], [19]. We present simulation results relaxing this assumption and discuss its implications in Sec. VIII-F. We also assume an ideal physical channel; that is, we ignore the possibility of preamble loss due to low signal power and omit the effects of power ramping as in [19]. We also assume that a collision is instantly detected and the colliding node

is ready to transmit a preamble in the next RAO. Further, we assume that there is no PDCCH congestion; that is, a successful preamble transmission always results in successful channel access. Despite these assumptions, the proposed expressions are still valid as a lower bound on the average access delay as the relaxation of these assumptions always results increased channel access delay. We also assume that there is no capture; that is, none of the colliding preambles can be detected. We relax this assumption in Sec. VII and derive an increased capacity expression which can easily be used in the proposed analysis.

IV. Optimum Load Controller

In this part, our aim is to quantify the delay bounds of a sub-optimal load controller by investigating an optimum load controller. The optimum load controller perfectly knows the total number of backlogged nodes, g, which includes both newly arriving nodes and the nodes failed previously. The controller announces an access probability q before each RAO.

For n nodes randomly selecting p preambles, the expected number of successful nodes can be written as :

$$u(n) = n\left(1 - \frac{1}{p}\right)^{n-1} \tag{1}$$

which is maximized at $n^* = p$; that is, the number of nodes transmitting a preamble at a RAO should be equal to the number of preambles. Hence, the base station should distribute the optimum access probability $q^*(g) = \min(1, p/g)$ to maximize channel utilization [7], [8], [34], [35].

Since this optimum controller maximizes the expected number of successful transmissions at a RAO, it also minimizes the overall channel access delay when a large number of devices arrive. The expected number of nodes successfully transmitting a preamble when this optimum controller is employed can be written as follows:

$$u^*(b) = \begin{cases} g\left(1 - \frac{1}{p}\right)^{g-1} & g < p \\ p\left(1 - \frac{1}{p}\right)^{p-1} \approx p\frac{1}{\mathrm{e}} & g \geq p \end{cases} \tag{2}$$

since $\left(1 - \frac{1}{p}\right)^{p-1} \approx \frac{1}{\mathrm{e}}$ where e is the Euler's number. This approximation is very accurate as the approximation error is less than 6% for $p \geq 9$ and less than 3% for $p \geq 18$. So, at each RAO, a maximum of p/e nodes can be successful on the average. Since the average time between RAOs is τ, the maximum rate of successful departures from the channel can be written as:

$$\delta = \frac{p}{\tau \mathrm{e}} \tag{3}$$

which we will refer as the *channel capacity* in the rest of the article.

V. Delay Analysis

Using the expressions derived above, we now analyze the delay for three different arrival distributions: a) simultaneous arrival of nodes at time $t = 0$, b) uniformly distributed arrivals over a fixed time interval $[0, T]$ and c) beta-distributed arrivals over the time interval $[0, T]$ representing a serious congestion.

A. Simultaneous Arrival of Nodes

We first investigate the case where all N nodes arrive at time $t_{arr} = 0$. Such simultaneous arrival of nodes is unlikely in practice; however, we obtain the upper bound of access delay for the optimum controller by analyzing this case.

In this case, p/e nodes depart at each RAO on the average as long as there are more than p backlogged nodes. Hence, the rate of successful departures equals to the channel capacity δ. Omitting the last few RAOs where the backlog reduces below p, the backlog is eliminated at time $t_{end} = \frac{N}{\delta}$ which results in the following expected departure time which also equals to the expected delay:

$$d_s = E[t_{dep}] = \frac{N}{2\delta}. \tag{4}$$

This expression implies that the access delay linearly decreases with the channel capacity. Hence, due to (3), the delay decreases linearly with the number of preambles and with the rate of RAOs.

B. Uniformly Distributed Traffic

In this part, we investigate the case where new nodes arrive to the channel at the following uniform rate:

$$\lambda(t) = \begin{cases} \frac{N}{T} & 0 < t < T, \\ 0 & o.w. \end{cases} \tag{5}$$

We investigate two different cases depending on rate of arrival: In the first case, $\lambda(t) < \delta$; i.e. the arrival rate is lower than the channel capacity so that the backlog is stable. In the second case, the arrival rate is greater than the channel capacity; i.e. $\lambda(t) > \delta$. In the latter case, the average number of nodes arriving between two RAOs is greater than p/e; so, the backlog increases as long as nodes continue to arrive the channel. The stability condition $\lambda(t) < \delta$ can also be rewritten as:

$$N < T\delta. \tag{6}$$

1) Stable Backlog: If the backlog is stable, the average number of successful departures at each RAO equals to the average number of arriving nodes between two RAOs. If the average number of nodes arriving between each RAO is x, the average number of backlogged nodes converges to $g(x) > x$ such that the average number of successful nodes at each RAO equals to x. That is, when all of the $g(x)$ backlogged nodes transmit a preamble, x of them will be successful on the average. If we substitute $g(x)$ in (2) as:

$$g(x)\left(1 - \frac{1}{p}\right)^{g(x)-1} = x, \tag{7}$$

we can obtain the average backlog as given by

$$g(x) = \frac{W\left(\frac{(p-1)x\log\left(1-\frac{1}{p}\right)}{p}\right)}{\log\left(1 - \frac{1}{p}\right)} \tag{8}$$

where $W(.)$ is the principal branch of the Lambert W function. Lambert W function is the inverse function associated with the

equation $We^W = f(W)$. It is encountered in different domains and its numerical solution is widely studied [36].

Approximating $\frac{p-1}{p} \approx 1$ and $\log(1-\frac{1}{p}) \approx -\frac{1}{p}$, $g(x)$ can be simplified as follows:

$$g(x) \approx -pW\left(-\frac{x}{p}\right). \tag{9}$$

For the arrival rate given by (5), the average number of nodes arriving between each RAO is $N\frac{\tau}{T}$ so the average backlog is $g(N\frac{\tau}{T})$. Since the number of backlogged nodes is always less than p, each node will transmit a preamble at each RAO and $N\frac{\tau}{T}$ of them will be successful on the average. Then, the success probability at each RAO is $\frac{N\frac{\tau}{T}}{g(N\frac{\tau}{T})}$. The expected delay can be found by multiplying the expected number of retransmissions with the duration between two RAOs as follows:

$$d_u = \left(\frac{g(N\frac{\tau}{T})}{N\frac{\tau}{T}} - 1\right)\tau, \quad \text{if } N < T\delta. \tag{10}$$

2) Unstable Backlog: In this case, the number of nodes arriving to the channel is higher than the channel capacity, i.e. $\lambda(t) > \delta$; so, the backlog increases continuously as long as the nodes continue to arrive. In this case, the departure behavior of nodes is similar to the simultaneous arrivals case since the backlog immediately becomes greater than the number of preambles. The optimum load controller will broadcast q^* such that the number of successful nodes at each RAO equals to p/e which corresponds a successful departure rate equal to the channel capacity, δ. Hence, the backlog will be eliminated at time $t_{end} = N/\delta$ which results in the expected departure time $E[t_{dep}] = \frac{N}{2\delta}$.

Unlike the simultaneous arrivals case, the mean arrival time of nodes is given by $E[t_{arr}] = T/2$ and the average delay experienced by a user can be written as:

$$d_u = E[t_{dep}] - E[t_{arr}] = \frac{N}{2\delta} - \frac{T}{2}, \quad \text{if } N > T\delta. \tag{11}$$

Similar to the simultaneous arrival of all users, the delay decreases linearly with the number of preambles and with the rate of RAOs for the unstable backlog case.

C. Beta-Distributed Traffic

3GPP suggests the use of the Beta distribution to model correlated M2M arrivals [4]; however, it has been acknowledged that the beta-distribution is hard to be analyzed mathematically [17]. Fortunately, similar difficulties regarding the Beta distribution had been encountered earlier in the risk analysis literature where the Beta distribution has been used to model uncertain durations. For risk analysis purposes, triangular approximations to the Beta distribution had been proposed [37]; and, we use these approximations to analyze the beta-distributed M2M traffic. The triangular distribution is more suitable for mathematical analysis than the Beta distribution and the differences between the two distributions do not create a significant difference in modeling the access delay of M2M nodes.

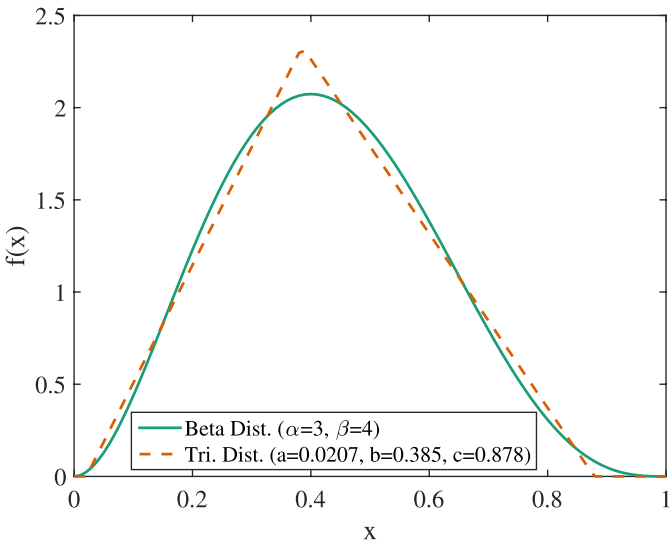

Fig. 2. Probability density function of the Beta distribution along with the density of its triangular approximation.

The density function of the Beta distribution for shape parameters α and β is given by

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{12}$$

where $B(\alpha,\beta)$ is the Beta function. When N nodes arrive over a time interval T according to the Beta distribution, the arrival rate of nodes can be written as [4]

$$\lambda(t;\alpha,\beta) = N\frac{t^{\alpha-1}(T-t)^{\beta-1}}{T^{\alpha+\beta-1}B(\alpha,\beta)}. \tag{13}$$

The triangular density function is also given by

$$h(x;a,b,m) = \begin{cases} \dfrac{2}{b-a}\dfrac{x-a}{m-a} & a \le x \le m, \\ \dfrac{2}{b-a}\dfrac{b-x}{b-m} & m \le x \le b. \end{cases} \tag{14}$$

Similarly, the arrival rate of nodes according to the triangular distribution can be written as

$$\lambda(t;a,b,m) = \begin{cases} \dfrac{N}{T}\dfrac{2}{b-a}\dfrac{t/T-a}{m-a} & a \le t/T \le m, \\ \dfrac{N}{T}\dfrac{b-t/T}{b-m} & m \le t/T \le b. \end{cases} \tag{15}$$

We use the following approximation which uses the 10% fractiles of the Beta distribution as given by

$$\begin{aligned} a &= 2.24Q_{0.1} - 1.63Q_{0.5} + 0.39Q_{0.9} \\ m &= -1.38Q_{0.1} + 3.78Q_{0.5} - 1.40Q_{0.9} \\ b &= -0.08Q_{0.1} - 0.72Q_{0.5} + 1.80Q_{0.9} \end{aligned} \tag{16}$$

where Q_u is the x such that $F(x) = u$ [37] where F is the CDF of beta distribution. An approximation of the Beta function with parameters $\alpha = 3$ and $\beta = 4$ along with its triangular approximation can be seen in Fig. 2.

In the rest of the paper, we use this approximation in the performance analysis of the beta-distributed traffic. We compare the performance of the proposed analysis against simulations

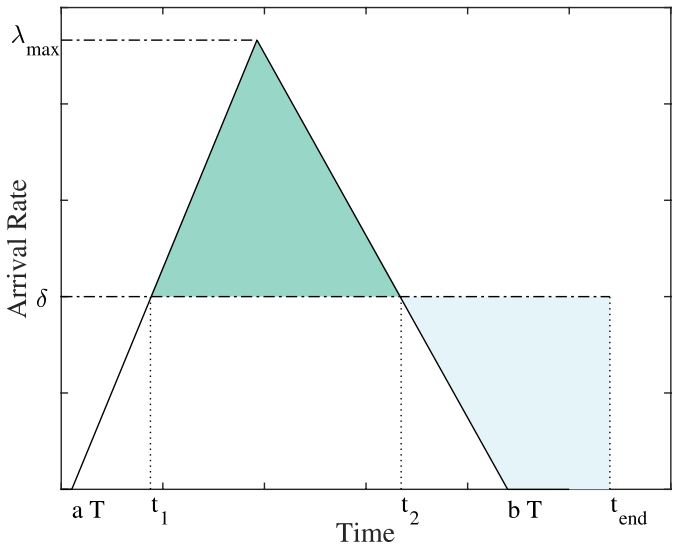

Fig. 3. The unstable backlog case where the maximum arrival rate significant exceeds the channel capacity such that the backlog accumulated between times t_1 and t_2 is eliminated after the end of nodes arrivals at time $t_{end} > bT$.

in which the nodes arrive according to a Beta distribution in Sec. VIII; and, the results suggest that the analysis based on the triangular distribution provides a very close estimation of access delay.

When the nodes arrive according to the triangular distribution (or according to the Beta distribution), the arrival rate is very low at the beginning. As time progresses, however, the arrival rate may exceed the channel capacity and the backlog may become unstable. We analyze the delay performance of the triangularly distributed traffic for two different cases. First, we analyze the stable case where the maximum arrival rate does not exceed the channel capacity. Second, we analyze the unstable case where the arrival rate significantly exceeds the channel capacity as shown in Fig. 3. In the figure, the arrival rate becomes larger than δ at time t_1 and decreases below δ at time t_2 and the backlog accumulated between times t_1 and t_2 is eliminated after the end of nodes arrivals at time $t_{end} > bT$. We exclude the analysis of the mildly congested case where the accumulated backlog is eliminated before the end of arrivals.

1) Stable Backlog: In this case, the maximum arrival rate never exceeds the channel capacity, $\lambda_{\max} < \delta$, which causes a stable backlog for the whole duration of arrivals. This condition can be written as follows:

$$N < \frac{1}{2}\delta T(b-a). \tag{17}$$

Here we assume that the channel is static for a node between its arrival and its successful departure; that is, the number of backlogged nodes is constant at the RAOs that the node transmits a preamble. Since the access delays are very small when the backlog is stable, this is a realistic assumption which is also confirmed by the accuracy of the simulation results. Using this assumption, the expected delay of a single node can be approximated by the delay obtained for the uniformly distributed traffic which is given by (10).

For the triangular arrival distribution, the maximum value of $\lambda(t)$ is $\lambda_{\max} = \frac{2}{b-a}\frac{N}{T}$; hence, the maximum number of nodes arriving between two RAOs can be written as $n_{\max} = \lambda_{\max}\tau$. Since the number of nodes arriving between two RAOs has a uniform distribution between 0 and $n_{\max}$ when the arrival distribution is triangular, the expected delay can be written as:

$$\begin{aligned} d_b &= \tau\left(\int_0^{n_{\max}} (\frac{g(n)}{n} - 1)n\mathrm{d}n\right) / \left(\int_0^{n_{\max}} n\mathrm{d}n\right) \\ &= \tau\left(-\frac{2}{k^2} - \frac{2W(-k)}{k} - \frac{2}{kW(-k)} + \frac{2}{k} - 1\right), \\ &\qquad \text{if } N < \frac{1}{2}\delta T(b-a). \end{aligned} \tag{18}$$

where $k = n_{\max}/p$.

At the limit of stable load case where $k = 1/\mathrm{e}$,

$$d_b = \tau(2\mathrm{e}(3-\mathrm{e}) - 1) \approx 0.53\tau, \quad \text{if } N = \frac{1}{2}\delta T(b-a) \tag{19}$$

i.e. average delay is approximately the half of the duration between two RAOs. So, it can be said that the delay experienced by nodes is insignificant if the arrival rate is below the channel capacity.

2) Unstable Backlog: In this case, the maximum arrival rate of nodes significantly exceeds the channel capacity such that the backlog accumulated during the peak of node arrivals is eliminated after the end of arrivals. As shown in Fig. 3, the backlog accumulated between times t_1 and t_2 starts to decrease at time t_2 until it is completely eliminated at time t_{end}.

We first obtain t_1 and t_2 which are the solutions of $\lambda(t) = \delta$ as follows:

$$t_1 = \frac{\delta T^2(a-b)(a-m)}{2N} + aT \tag{20}$$

$$t_2 = \frac{\delta T^2(a-b)(b-m)}{2N} + bT. \tag{21}$$

The area of the heavily shaded region in Fig. 3 gives the number of accumulated backlogged nodes between times t_1 and t_2 as the integral of the arrival rate gives the number of nodes. Similarly, the area of the lightly shaded area gives amount of departures from this accumulated backlog after time t_2. The time at which the backlog is completely eliminated, t_{end}, can be found by equating the heavily shaded area with the lightly shaded area:

$$\frac{1}{2}(\lambda_{max} - \delta)(t_2 - t_1) = \frac{1}{2}(bT - t_2)\delta + (t_{end} - bT)\delta \tag{22}$$

which results in the following solution:

$$t_{end} = \frac{\delta T^2(a-b)(a-m)}{4N} + aT + \frac{N}{\delta}. \tag{23}$$

This analysis is valid for $t_{end} > bT$ which can be rewritten as:

$$N > \frac{1}{2}\delta T\left(\sqrt{(b-a)(b-m)} - a + b\right). \tag{24}$$

We compute the mean delay experienced by the nodes arriving after time t_1 and assume that the delay experienced by the nodes arriving before time t_1 is zero for the sake of simplicity. We believe this assumption is reasonable because, as given by (19), the access delays are negligible when the backlog is stable. Besides, the numerical results given in Sec. VIII suggest that this assumption does not significantly affect the accuracy of analysis.

The mean arrival time of nodes arriving after time t_1 can be found as:

$$\begin{aligned} E[t_{arr}] &= \left(\int_{t_1}^{T} \lambda(t) t \mathrm{d}t\right) / \left(\int_{t_1}^{T} \lambda(t) \mathrm{d}t\right) \\ &= \Big(-\delta^3 T^4 (a-b)^2 (a-m)^2 - 3a\delta^2 N T^3 (a-b) \\ &\quad \times (a-m) + 4N^3 T (a+b+m)\Big) \\ &\quad / \Big(12N^3 - 3\delta^2 N T^2 (a-b)(a-m)\Big). \end{aligned} \tag{25}$$

The successful departure rate of nodes arriving after time t_1 equals to δ. Then, the mean departure time of nodes arriving after time t_1 is given by:

$$\begin{aligned} E[t_{dep}] = (t_{end} - t_1)/2 + t_1 = \frac{3\delta T^2 (a-b)(a-m)}{8N} \\ + aT + \frac{N}{2\delta}. \end{aligned} \tag{26}$$

Then, by assuming that the access delay of nodes arriving before time t_1 as negligible, the mean delay can be written as follows:

$$\begin{aligned} d_b &= (E[t_{dep}] - E[t_{arr}]) N_{[t_1, bT]} / N \\ &= \Big(32\delta N^3 T (2a - b - m) + 24\delta^2 N^2 T^2 (a-b)(a-m) \\ &\quad - \delta^4 T^4 (a-b)^2 (a-m)^2 + 48N^4\Big) / \Big(96\delta N^3\Big), \\ &\quad \text{if } N > \frac{1}{2}\delta T \left(\sqrt{(b-a)(b-m)} - a + b\right). \end{aligned} \tag{27}$$

where $N_{[t_1, bT]}$ is the number of nodes arriving after time t_1.

VI. Estimation Error on the Channel Capacity

So far, we have assumed that the optimum load controller has the perfect information about the number of backlogged nodes in the network. If the base station makes a proportional error of r such that is estimates the backlog as gr, it distributes an erred access probability of $\min(1, p/gr)$. If the base station estimates that there are less than p backlogged users, i.e. $gr < p$, all of the g nodes will transmit a preamble since the base station will broadcast $q = 1$. If the base station estimates that the number of nodes is greater than p, the base station will distribute $q = p/gr$ and will result in p/r nodes to transmit a preamble. By modifying (2), the average number of nodes which will be successful at a RAO when there is an estimation error can be written as:

$$u^{err}(g, r) = \begin{cases} g\left(1 - \frac{1}{p}\right)^{g-1} & \text{if } gr < p \\ \frac{p}{r}\left(1 - \frac{1}{p}\right)^{\frac{p}{r}-1} & \text{if } gr \geq p. \end{cases} \tag{28}$$

Then, the ratio of maximum number of successful nodes when there is an estimation error to the maximum number of successful nodes when there is no estimation error can be found by dividing (28) by (2) as:

$$\frac{u^{err}(g, r)}{u^*(g, r)} = \begin{cases} 1 & \text{if } gr < p \text{ and } g < p \\ \frac{g}{p}\left(1 - \frac{1}{p}\right)^{g-p} & \text{if } gr < p \text{ and } g \geq p \\ \frac{1}{r}\left(1 - \frac{1}{p}\right)^{\frac{p}{r}-p} & \text{if } gr \geq p \text{ and } g \geq p \end{cases} \tag{29}$$

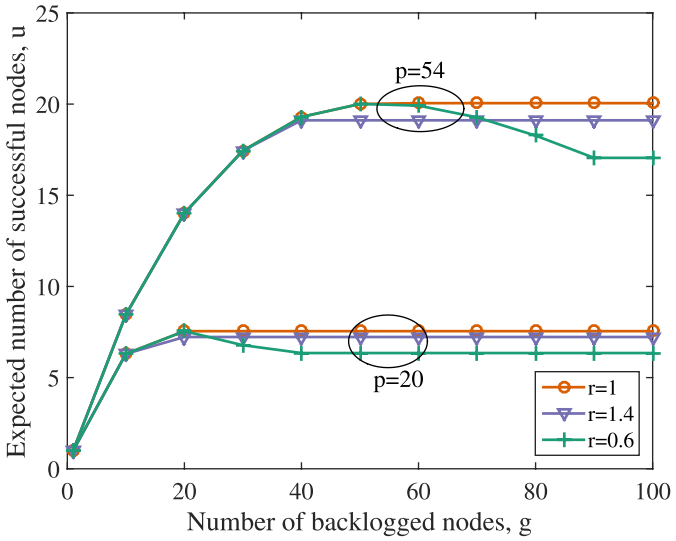

Fig. 4. The expected number of successful nodes for different estimation error ratios as the number of backlogged nodes changes.

Fig. 4 plots u for $r = 1$, $r = 0.6$ and $r = 1.4$ as g increases for $p = 54$ and $p = 20$. For low g, it can be seen that the estimation error does not change utilization since the base station distributes an access probability of $q = 1$ even if there is a significant estimation error. As g increases, however, the estimation error starts to reduce the expected number of successful nodes. However, this reduction is not significant as the utilization function is a smooth function around its maximum point. For example, for $p = 54$, the utilization reduces at most 15% when $r = 0.6$, i.e. there is a 40% underestimation. If there is a 40% overestimation, the maximum reduction in the utilization is merely 5%. It can be said that overestimation is better than underestimation, so the estimators in practical load control algorithms should be more pessimistic about the number of backlogged nodes rather than being optimistic.

At the first sight, it can be inferred that the estimation accuracy is not crucial as the reduction in the system capacity is insignificant even if the estimation error is very large. However, the accuracy becomes important if the reduced system capacity causes the channel to become unstable. Numerical results presented in Sec. VIII-C show that the estimation error becomes crucial when the system is operating close its stability limits. In that case, a reduction in the capacity may cause an accumulation of backlog resulting in a severe access delay.

It should be noted that the estimation errors in practical systems may not be in the form of a proportional error; but, we leave further study of more complex error models as a future work.

VII. Channel Capacity Under Perfect Capture

Up to this point, we have assumed that none of colliding preambles resulted in successful channel access. In practice, however, one of the interfering transmissions can be detected if the SNR is sufficiently high. For the perfect capture case,

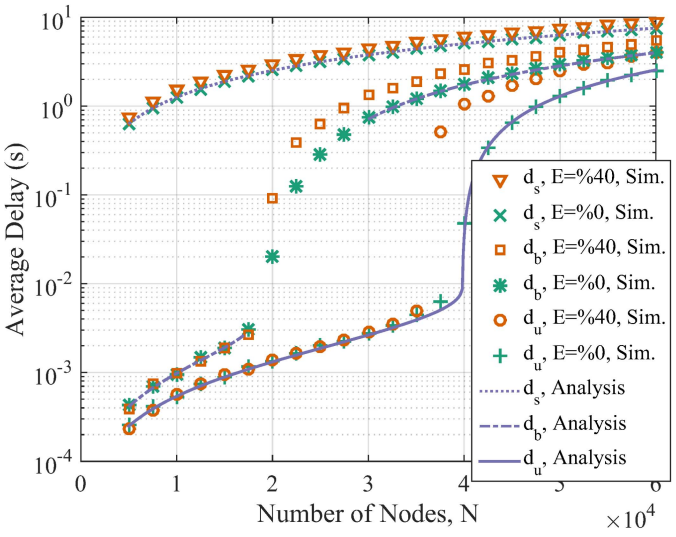

Fig. 5. Average delay as N increases for different arrival distributions and estimation errors.

in which one of the colliding preambles are detected with probability one, the preambles chosen by at least one node results in successful channel access. Then, for the optimal controller, the maximum channel utilization can be written as

$$u^{cap} = p - p\left(1 - \frac{1}{p}\right)^{p} \approx p\left(1 - \frac{1}{\mathrm{e}}\right), \quad \text{if } g \geq p. \tag{30}$$

as the probability of a preamble not selected by any of the p nodes can be written as $(1 - 1/p)^p$. Then, the channel capacity is increased to $\delta^{cap} = \frac{p(1-1/\mathrm{e})}{\tau}$. It is possible to use this new capacity expression in the proposed delay analysis to obtain the delay in the case of perfect capture. It should also be noted that the load controller can further be optimized for a given probability of capture but we leave it as a future study.

VIII. Numerical Results

We evaluated the proposed bounds using an LTE random access simulator that we developed using MATLAB. We performed simulations for beta-distributed arrivals, uniformly distributed arrivals and for the simultaneous arrival of nodes by varying N between 5000 and 60 000 for $T = 10$ s. These scenarios have been used in the numerical evaluation of various load control algorithms in various studies and 3GPP documents [4], [9], [10], [24]. For beta-distributed arrivals, we selected the shape parameters as $\alpha = 3$ and $\beta = 4$ similar to [4]. We have simulated a simplified version of the random access procedure with the assumptions of perfect physical layer, no capture effect and no PDCCH congestion as described in Sec. III.

Fig. 5 plots the change in the average delay as N increases for different estimation accuracies about the number of backlogged nodes. The estimate at the base station is obtained by multiplying the actual number of backlogged nodes by a random factor. For example, to obtain a $E = 40\%$ estimation error, the actual number of nodes is multiplied by a uniformly distributed random number between 0.2 and 1.8 such that the mean estimation error is 40 %. The estimation is computed independently for each RAO and the base station broadcasts the access probability based on this estimation. $E = 0\%$ case implies that the base station has perfect information about the number of backlogged nodes. For this plot, $p = 54$ preambles are allocated for M2M traffic similar to previous studies [4] and the RAOs are presented once in $\tau = 5$ ms as defined by the PRACH Configuration Index 6 [38].

For the uniformly distributed traffic, Fig. 5 plots the proposed analysis given by (10) and (11) for the stable and the unstable backlog case, respectively. Analysis of the stable case is valid for $N < 39\,731$ as found by (6). For the beta-distributed traffic, the figure plots the proposed delay expressions given by (18) and (27) for the stable and the unstable case, respectively. The parameters of the triangular approximation is $a = 0.0208$, $b = 0.879$, $m = 0.385$ according to (16). Stable backlog analysis is valid for $N < 17\,043$ as given by (17) and the unstable backlog analysis is valid for $N > 29\,968$ as given by (24). We note that the proposed bounds does not cover the mildly congested case which is between $17\,043 < N < 29\,968$ for this set of parameters. For the simultaneous arrivals case, the figure plots the proposed delay expression given by (4).

A. Accuracy of the Proposed Bounds

The proposed analyses predict the delay very accurately for all traffic distributions for $E = 0\%$. The average error is less than 1 % for the simultaneous arrival of nodes, less than 3 % for the beta-distributed traffic and less than 6 % for the uniformly distributed traffic. These results also suggest that the triangular distribution provides a good approximation for the analysis of the Beta distribution as far as the M2M traffic is considered.

B. Effect of the Arrival Distribution

According to (6) and (17), when the arrivals are beta-distributed, the number of nodes that can stably be supported over a fixed time interval is $\frac{b-a}{2} = 0.43$ times the number of nodes that can be stably supported when the arrivals are uniformly distributed. For the case we considered, uniformly-distributed arrival of 39 000 nodes can stably be supported by the random access channel but only 17 000 nodes can be supported when the arrivals are beta-distributed. Hence the network has to be significantly over-provisioned for the case that the arrivals are strongly syncronized.

Fig. 6 shows the ratio of average delay when the arrivals are simultaneous, d^s, and beta-distributed, d_b, to the delay when the arrivals are uniformly distributed, d_u. For small N, the backlog stays low for both the uniformly distributed and the beta-distributed traffic so their delay behavior is similar. For large N, the delay of all distributions approach each other since the backlog is unstable regardless of the arrival distribution. The difference between the Beta and uniform distributions is more apparent between $N = 17\,043$ and $39\,731$ where the backlog is stable for the uniformly distributed traffic but unstable for the beta-distributed traffic. For $N = 32\,500$, for example, the mean delay is approximately 300 times more for the beta-distributed traffic in comparison to the uniformly distributed traffic for $E = 0\%$.

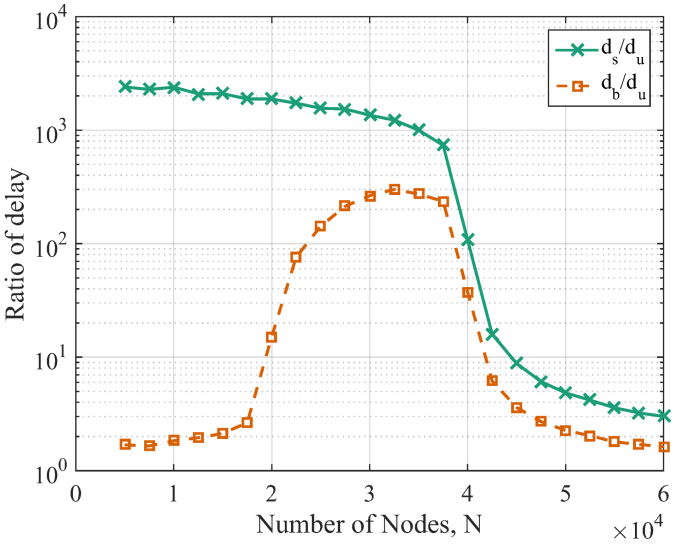

Fig. 6. The ratio of average delay for the simultaneous arrivals and for the beta-distributed arrivals to the average delay for the uniformly-distributed arrivals.

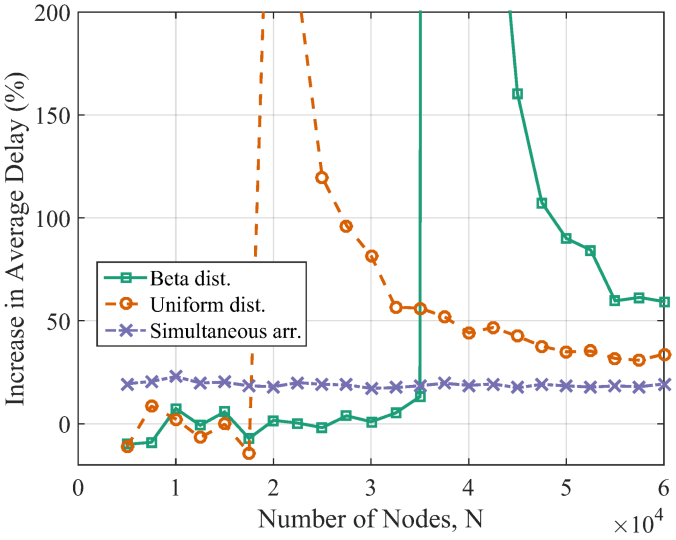

Fig. 7. Increase in the average delay when there is a 40% average estimation error.

Simultaneous arrivals set the upper bound of delay for a given number of nodes. Even for low N, the delay is significantly higher than the uniformly distributed arrivals. As N increases, the delay of both the uniformly-distributed traffic and the beta-distributed traffic approaches to the delay of the simultaneous arrivals case.

C. Effect of the Estimation Error

Fig. 7 shows the percentage increase in delay when $E = 40\%$ as N increases for all traffic distributions. For the simultaneous arrival of all users, the estimation error causes a 20 % increase in delay independent of the number of nodes. Since the channel is overloaded for all N, the increase in delay is proportional to the reduction in the channel capacity and is independent from the number of nodes.

For the uniformly distributed traffic, the average delay does not increase for $E = 40\%$ for $N \leq 35\,000$. For small N, the estimation error does not cause an increase in delay because the base station broadcasts an access probability of one even if there is a significant estimation error. For example, if the backlog is estimated as 10 instead of 5 with a 100 % error, the base station still broadcasts $q^* = 1$ which is the optimum q value for both $g = 5$ and $g = 10$.

For $N = 37\,500$, close to the stability threshold of $N = 39\,731$, the delay for $E = 40\%$ is 80 times more than the delay for $E = 0\%$ because the estimation error results a suboptimal utilization and causes an unstable backlog. This suggests that the estimation accuracy is very crucial if the channel load is close to its capacity limits. After this point, the delay increase due to the estimation error starts to reduce as N increases and approach to the delay increase for simultaneous arrival of all users.

A similar behavior can be observed for the beta-distributed traffic: The delay does not increase due to estimation error for $N \leq 15\,000$. Close to the stability limit, at $N = 20\,000$, the delay is 3.5 times larger for $E = 40\%$. Such a sharp increase occurs when the channel is close to its stability limit. As N increases, the increase in delay due to the estimation error starts to reduce.

From these results it can be inferred that the estimation accuracy is not crucial both when the system is lightly loaded and when the system is heavily overloaded. A pessimistic estimation error of 40 % results in a 20 % increase in delay when the system is heavily overloaded. There is no increase in delay when the system is not lightly loaded because all nodes should transmit a preamble. Close to the stability limit, however, the estimation accuracy becomes crucial since the reduction in the channel capacity may cause the backlog to be unstable.

D. Effect of the Frequency of RAOs

There are different LTE random access configurations which define different periods for RAOs ranging from 1 ms to 20 ms [2], [38]. Fig. 8 plots the average delay for different random access periods, τ, for $E = 0\%$ along with the proposed bounds. The proposed bounds estimate the delay with less than 4 % average error for both traffic distributions. Fig. 9 plots the change in delay as the RAO period increases for the uniformly-distributed arrivals. The results suggest that it is possible to reduce delay several orders of magnitude by increasing the frequency of RAOs. For $N = 40\,000$, the average delay reduces to 2.6 ms from 1 s when τ is reduced from 6 ms to 4 ms for the uniformly distributed arrivals. This dramatic reduction occurs since the backlog is stable for $\tau = 4$ ms and unstable for $\tau = 6$ ms. Similar significant changes in delay can be observed for different N. Hence, it can be said that significant improvements in delay can be obtained if the channel can be stabilized by increasing the frequency of RAOs.

E. Effect of the Number of Preambles Reserved for M2M

There are proposals for splitting the set of preambles for M2M and H2H traffic [28] which can be used adapt to

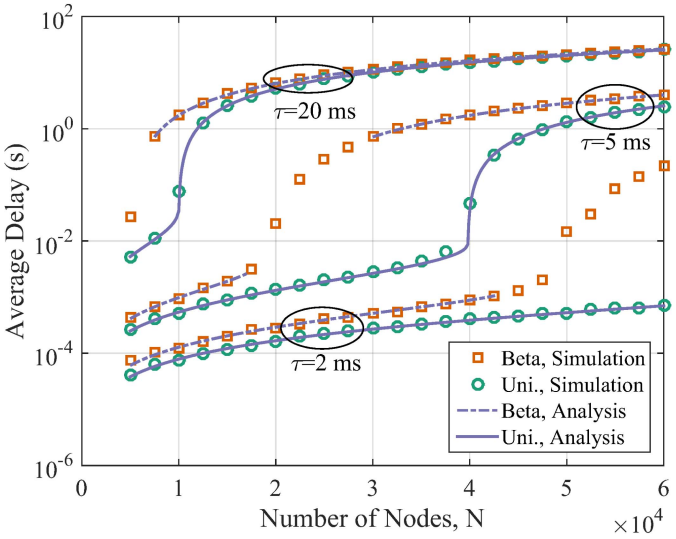

Fig. 8. Average delay as N increases for different RAO periods.

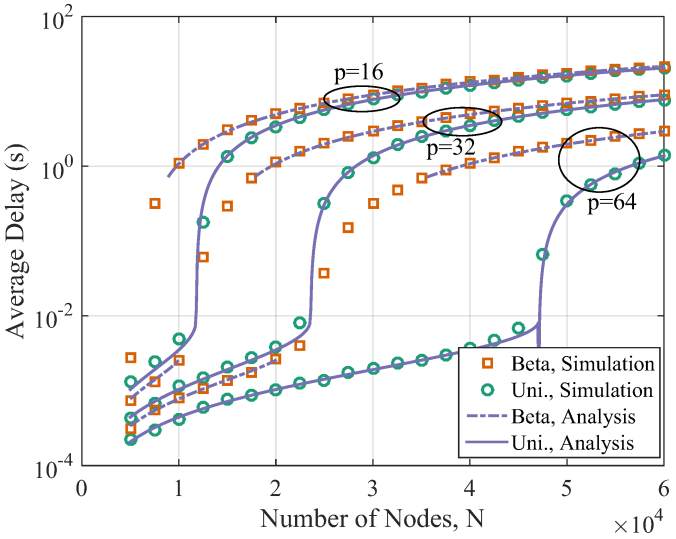

Fig. 10. Average delay as N increases for different number of preambles.

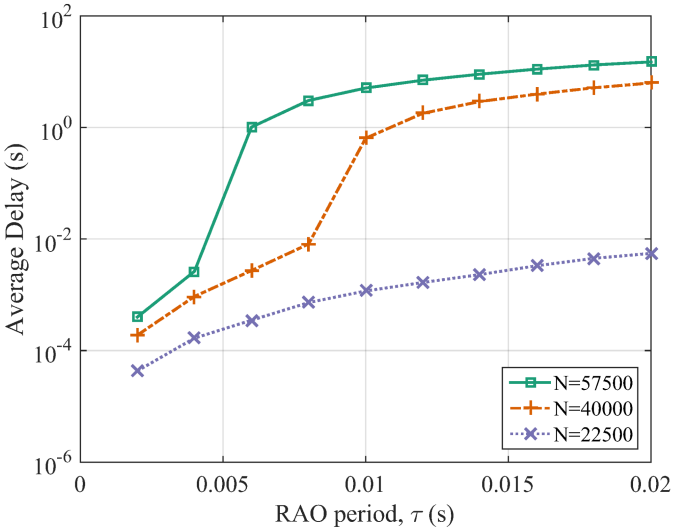

Fig. 9. Average delay as RAO period increases for different N.

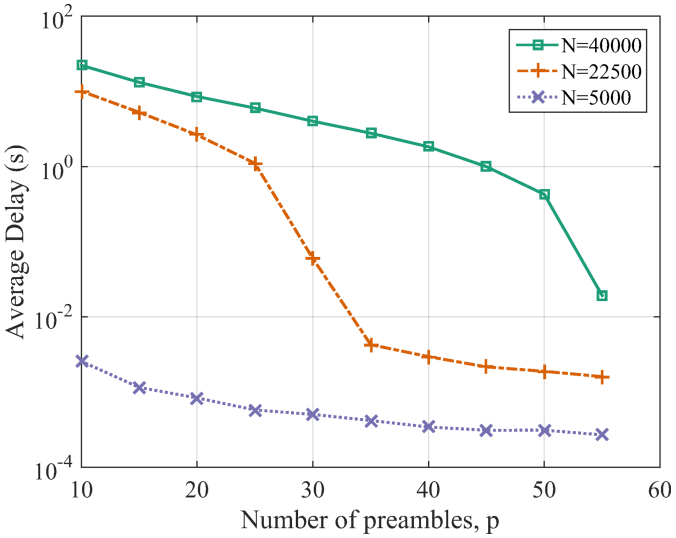

Fig. 11. Average delay as number of preambles reserved for M2M traffic increases for different N.

control the delay experienced by M2M or human devices. For that reason, we analyzed the change in the delay for different number of preambles allocated for M2M traffic, p, for $\tau = 5$ ms and $E = 0\%$ (Fig. 10). For the simulated points, the proposed analysis estimates the delay with less than 6 % average error. As it can be seen from Fig. 11, by increasing p, it is possible to obtain dramatic improvements in delay similar to the previous case. For example, for $N = 22\,500$, the delay reduces from 1.1 s to 4.2 ms when the p is increased from 25 to 35. Since both p and $1/\tau$ affects the channel capacity in the same manner as it can be seen in (3), both system parameters can be tuned to adapt the access delay.

F. Effect of Backoff Parameters

Our analysis assumes that the nodes remain backlogged if their preamble transmissions are unsuccessful and there is no limit on the number of retransmissions. In practice, however, the nodes typically backoff after an unsuccessful preamble transmission and give up random access after reaching a limit on the retransmissions [20], [39], [40]. In the LTE standard, the backoff duration is determined by the eNodeB and broadcasted to all nodes.

To evaluate performance of the proposed delay bounds when there is a random backoff, we performed simulations for beta-distributed arrival of 10 000 and 30 000 nodes over a 10 s interval for $p = 54$ and $\tau = 5$ ms. In these simulations, the nodes backoff for a random number of RAOs uniformly distributed between 0 and U both when their preamble transmissions are unsuccessful and when they are barred due to the access probability distributed by the base station. We denote the maximum allowable number of backoffs as W.

For lower values of W, most of the nodes give up channel access in the case of a congestion which we call as *outage*. This effect results in skewed statistics of access delay because the nodes which experienced the largest amount of delay gave

TABLE I
CHANGE IN DELAY FOR DIFFERENT BACKOFF PARAMETERS

N=10 000			N=30 000		
W	U	Delay	W	U	Delay
5	5	2.835 ms	5	788	1.314 s
10	4	2.871 ms	10	298	0.955 s
15	3	2.448 ms	15	182	0.873 s
20	2	1.968 ms	20	130	0.834 s
35	1	1.484 ms	35	69	0.778 s
50	0	1.000 ms	50	46	0.755 s
∞	0	1.000 ms	∞	0	0.736 s
Bound (18)		0.979 ms	Bound (27)		0.712 s

up before accessing the channel. To perform a fair comparison, we constrained the outage probability to 1% and obtained the minimum U values which satisfies this constraint for a set of different W values.

Table I shows how the average delay changes for various values of W and U representing the averages of 1000 simulations. For lower values of W, the backoff duration has to be high to satisfy the outage constraint. Due to the increasing backoff durations, the average delay experienced by the nodes increases as W reduces; and, the delay approaches to the lower bound as W increases.

It should be noted that the number of attempting nodes are still controlled by the optimum controller even if there is a backoff procedure; that is, a node can still be barred by the base station after becoming active following a backoff. Hence, the number of collisions do not increase even for lower values of U.

For $N = 10\,000$, the channel is lightly loaded and most nodes become successful on their first attempts resulting in a very low access delay around 1 ms. For $N = 30\,000$, however, the channel is seriously loaded leading to access delays more than 0.7 s.

We note that the proposed analysis still provides a lower bound of delay when there is a backoff mechanism. In our simulations, the maximum backoff duration is kept fixed for the duration of the simulations. In practice, however, it is possible to optimize the backoff duration depending on the network load. Such an optimization is beyond the scope of this paper; but, the proposed bounds can also be used as a benchmark for studies on backoff parameter optimization as they still define the achievable lowest average delay.

IX. CONCLUSIONS

We proposed random access delay bounds for massive M2M traffic carried over LTE networks for both uniformly distributed and the beta-distributed M2M arrivals. Through the help of a triangular approximation, we obtained closed-form expression of average delay for beta-distributed arrivals for the first time to the best of our knowledge. Numerical results show that the proposed bounds model the delay very accurately for both distributions as the maximum average error is less than 6%.

Our results show that the channel should be significantly over-provisioned for the case of correlated M2M arrivals modeled by a Beta distribution in comparison to the uniformly distributed arrivals. Such correlated arrivals are likely to occur in M2M scenarios due to some external triggering event; and, our results indicate that the average delay experienced by beta-distributed M2M arrivals can be 300 times more than the average delay of uniformly distributed arrivals.

Our results also indicate that the estimation accuracy of the channel backlog is not crucial when the network is either lightly loaded or heavily overloaded; but, such an error may increase access delay as much as 80 times if the channel is operating close to its capacity limits. Besides, we show that overestimating the number of nodes is better than underestimation at the same error levels; so, practical load controllers should tend be pessimistic about the number of nodes.

We have also shown that the access delay can be reduced by several orders of magnitude if the frequency of random access opportunities and the number of preambles for M2M communication is tuned appropriately when the random access channel becomes unstable.

Moreover, we have derived the capacity of the optimum controller in the case of perfect capture where one of the colliding preambles can always be detected. This expression can be used to compute the best-case delay where the receiver can distinguish among the collided preambles.

We have also considered the effect of backoff parameters and limits on the average delay. Our results show that the average delay reduces as the maximum allowable number of backoffs increases and approaches to the proposed bounds. Hence, our results can also be used as a benchmark in future studies on backoff parameter optimization.

ACKNOWLEDGMENT

The author is grateful to Ali T. Koc from Intel Corporation for introducing this topic to him.

REFERENCES

[1] OECD. "Machine-to-machine communications: Connecting billions of devices," OECD Publishing, Paris, France, OECD Digital Economy Papers 192, 2012. [Online]. Available: http://dx.doi.org/10.1787/5k9gsh2gp043-en

[2] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, First Quarter 2014. [Online]. Available: http://dx.doi.org/10.1109/SURV.2013.111313.00244

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

[4] 3GPP, "Study on RAN improvements for machine-type communications," 3GPP (3rd Generation Partnership Project), Sophia Antipolis Cedex, France, Tech. Rep. TR 37.868 V11.0.0, Sep. 2011.

[5] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: Overload control," *IEEE Netw.*, vol. 26, no. 6, pp. 54–60, Nov./Dec. 2012.

[6] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[7] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Stabilizing multi-channel slotted Aloha for machine-type communications," in *Proc. IEEE ISIT*, Jul. 2013, pp. 2119–2123.

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

[9] O. Arouk and A. Ksentini, "Multi-channel slotted Aloha optimization for machine-type-communication," in *Proc. ACM MSWIM*, 2014, pp. 119–125.

[10] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 5815–5820.
[11] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.
[12] 3GPP, "Service requirements for machine-type communications (MTC)," 3GPP (3rd Generation Partnership Project), Sophia Antipolis Cedex, France, Tech. Rep. TS 22.368 V13.1.0, Dec. 2014.
[13] 3GPP, "Study on machine-type communications (MTC) and other mobile data applications communications enhancements," 3GPP (3rd Generation Partnership Project), Sophia Antipolis Cedex, France, Tech. Rep. TR 23.887 V12.0.0, Dec. 2013.
[14] I. Vukovic and I. Filipovich, "Throughput analysis of TDD LTE random access channel," in *Proc. IEEE PIMRC*, Sep. 2011, pp. 1652–1656.
[15] R.-G. Cheng, C.-H. Wei, S.-L. Tsao, and F.-C. Ren, "RACH collision probability for machine-type communications," in *Proc. IEEE 75th Veh. Technol. Conf. (VTC Spring)*, May 2012, pp. 1–5.
[16] C. Wei, R.-G. Cheng, and S. Tsao, "Modeling and estimation of one-shot random access for finite-user multichannel slotted ALOHA systems," *IEEE Commun. Lett.*, vol. 16, no. 8, pp. 1196–1199, Aug. 2012.
[17] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy, "Impact of machine-type communications on energy and delay performance of random access channel in LTE-advanced," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 366–377, Jun. 2013.
[18] A. Mutairi, S. Roy, and G. Hwang, "Delay analysis of OFDMA-Aloha," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 89–99, Jan. 2013.
[19] A. M. Ahmadian, O. Galinina, S. Andreev, and Y. Koucheryavy, "Modeling contention-based M2M transmissions over 3GPP LTE cellular networks," in *Proc. IEEE ICC Workshops*, Jun. 2014, pp. 441–447.
[20] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, 2015, pp. 1–6.
[21] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.
[22] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb. 2015.
[23] 3GPP, "Access class barring and overload protection (release 7)," 3GPP (3rd Generation Partnership Project), Sophia Antipolis Cedex, France, Tech. Rep. TR 23.898 V7.0.0, Mar. 2005.
[24] U. Phuyal, A. T. Koc, M.-H. Fong, and R. Vannithamby, "Controlling access overload and signaling congestion in M2M networks," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2012, pp. 591–595.
[25] R.-G. Cheng, J. Chen, D.-W. Chen, and C.-H. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956–2968, Jun. 2015.
[26] J.-P. Cheng, C. H. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2011, pp. 368–372.
[27] M. K. Giluka, A. Prasannakumar, N. Rajoria, and B. R. Tamma, "Adaptive RACH congestion management to support M2M communication in 4G LTE networks," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2013, pp. 1–6.
[28] K.-D. Lee, S. Kim, and B. Yi, "Throughput comparison of random access methods for M2M service over LTE networks," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2011, pp. 373–377.
[29] D. Kim, W. Kim, and S. An, "Adaptive random access preamble split in LTE," in *Proc. 9th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jul. 2013, pp. 814–819.
[30] N. K. Pratas, H. Thomsen, Č. Stefanović, and P. Popovski, "Code-expanded random access for machine-type communications," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2012, pp. 1681–1686.
[31] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.
[32] E. Dahlman, S. Parkvall, and J. Skold, "Chapter 14—Access procedures," in *4G LTE/LTE-Advanced for Mobile Broadband*. Oxford, U.K.: Academic, 2011, pp. 301–321. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-385489-6.00014-X
[33] N. Zhang, G. Kang, J. Wang, Y. Guo, and F. Labeau, "Resource allocation in a new random access for M2M communications," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 843–846, May 2015.
[34] D. Shen and V. O. K. Li, "Performance analysis for a stabilized multi-channel slotted ALOHA algorithm," in *Proc. IEEE PIMRC*, Sep. 2003, pp. 249–253.
[35] P.-L. Chen, H.-P. Ho, C.-H. Chang, and H.-Y. Hsieh, "Analyzing and minimizing random access delay for delay-sensitive machine-to-machine communications: A new perspective on adaptive persistence control," in *Proc. IEEE Int. Conf. Internet Things*, Sep. 2014, pp. 69–74.
[36] E. W. Weisstein. (2002). *Lambert W-Function*, MathWorld. [Online] Available: http://mathworld.wolfram.com/LambertW-Function.html
[37] D. Johnson, "The triangular distribution as a proxy for the beta distribution in risk analysis," *J. Roy. Statist. Soc. D (Statistician)*, vol. 46, no. 3, pp. 387–398, 1997.
[38] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); physical layer procedures," 3GPP (3rd Generation Partnership Project), Sophia Antipolis Cedex, France, Tech. Rep. TS 36.213 V13.1.1, Mar. 2016.
[39] J.-B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.
[40] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č. Stefanović, and P. Popovski, "Assessment of LTE wireless access for monitoring of energy distribution in the smart grid," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 675–688, Mar. 2016, doi: 10.1109/JSAC.2016.2525639.

Mehmet Koseoglu (M'15) received the B.Sc., M.Sc., and Ph.D. degrees from Bilkent University, Ankara, Turkey, in 2004, 2007, and 2013, respectively, all in electrical and electronics engineering. From 2004 to 2006, he was a Software Engineer with Aselsan Inc., Ankara. From 2007 to 2013, he was a Research and Teaching Assistant with the Electrical and Electronics Engineering Department, Bilkent University. He is currently an Assistant Professor with the Department of Computer Engineering, Hacettepe University, Ankara. His current research interests are on M2M communications for 4G/5G systems and underwater acoustical networking.