Modeling and Analysis of Random Access Channels With Bursty Arrivals in OFDMA Wireless Networks

Chia-Hung Wei, Giuseppe Bianchi, and Ray-Guang Cheng, *Senior Member, IEEE*

***Abstract*—Random access channels (RACHs) in cellular networks are normally designed for Poisson-distributed arrivals with a constant rate. Unexpected bursty arrivals may result in severe collisions in RACHs and thus degrade users' service qualities. This paper presents an analytical model for investigating the transient behavior of the RACHs with bursty arrivals generated in a specific time interval for orthogonal frequency-division multiple access (OFDMA) wireless networks. The proposed model has considered the implementation details of the OFDMA random access procedure (such as periodic access characteristic, uniform random backoff policy, and power-ramping effect) and the effect of new arrivals. The performance metrics of collision probability, success probability, and average access delay and the cumulative distribution function of the number of preamble transmissions and access delay for the successfully accessed mobile station are then derived based on the analytical model. The accuracy of the proposed analytical model was verified through computer simulations, and the results show the effectiveness of the proposed model.**

***Index Terms*—Multichannel slotted ALOHA, random access, transient behavior.**

I. Introduction

In modern cellular networks, Mobile Stations (MSs) use Random-Access CHannels (RACHs) to perform initial association to the network, to request transmission resources, and to re-establish a connection to the Base Station (BS) upon failure. The random access procedure normally involves a four-message handshake between the MS and the BS. The four messages include a *Preamble Transmission* (or, Message 1 in Long Term Evolution (LTE)/LTE-Advanced (LTE-A) [1]) initiated by an MS; a *Random Access Response (RAR)* (or, Message 2) replied by the BS; a *Connection Request* (or, Message 3 (Msg3)) transmitted by the MS; and a *Connection Response* (or, Message 4 (Msg4)) confirmed by the BS [2]. The first two messages are exchanged over a shared transmission resource, whereas the remaining two messages are exchanged in dedicated logical channels specifically reserved for the MS. A random access attempt is completed if the four messages are successfully exchanged. For LTE, successful completion of a random access procedure implies the successful reception of Msg4. For Universal Mobile Telecommunications System (UMTS), completion of a random access procedure relies on the successful exchange of RACH messages [3].

The contention-based operation of the RACH is based on (multi-channel) ALOHA-type access, where each MS transmits its preamble in the first available random-access slot [2]. The design of the random-access schemes is always targeted for effective usage of the RACH (i.e., the successful transmissions of the preamble and message from the attempting MSs) [4], and has been largely investigated in the past. Analytical and simulation results have provided guidelines to optimize the RACH parameters for maximizing the access success probability [5], [6], reducing the collision probability [5], [7], and minimizing the access delay [4]–[6], [8]–[10]. In most studies, the authors assumed that new and backlogged arrivals follow a Poisson [4], [6], [8]–[12] or Bernoulli distribution [5] with a constant rate, and focused on the steady-state behavior of the RACHs. Several algorithms were developed to stabilize the RACH by throttling the arrival rate [10], [11] or adjusting the backoff parameter [4], [5], [9], [12] based on a fixed arrival rate and/or a constant successful transmission probability.

The emergence of Machine-to-Machine communication (M2M) brings about new requirements in the performance understanding of the RACH operation. M2M is an important service defined to facilitate machines communicating with each other over the next generation Orthogonal Frequency Division Multiple Access (OFDMA) cellular networks [4]. Even if the bandwidth requirement of M2M traffic may be limited, M2M services may involve tens of thousands devices per cell [3] which need to simultaneously reserve wireless access resources. Hence, the RACH may become the performance bottleneck of the wireless access network. Indeed, congestion in RACHs may block most of the random access attempts from MSs even if the network has lots of unused capacity and thus, may lead to an under-utilized network.

Moreover, a considerable amount of M2M applications are characterized by bulk arrivals and require very high energy efficiency to ensure the long lifetime of the network [2]. The traffic generated by the mass machine devices may generate bursty arrivals in a very short period of time. The bursty arrivals may cause the congestion at the RACHs and thus, result in intolerable delays, packet loss, or even service unavailability for existing Human-to-Human (H2H) services. This contrasts with most of the existing studies on multi-channel slotted ALOHA systems or RACHs in OFDMA cellular networks, focused on light-loaded H2H services and the steady-state performance of

Manuscript received February 25, 2014; revised July 22, 2014; accepted November 16, 2014. Date of publication December 4, 2014; date of current version April 7, 2015. This work was supported in part by the Ministry of Science and Technology of Taiwan under Contract MOST 102-2221-E-011-003-MY3. The associate editor coordinating the review of this paper and approving it for publication was A. A. Abouzeid. *(Corresponding author: Ray-Guang Cheng.)*

C.-H. Wei and R.-G. Cheng are with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: crg@ieee.org).

G. Bianchi is with the School of Engineering, University of Roma Tor Vergata, Rome 00173, Italy (e-mail: giuseppe.bianchi@uniroma2.it).

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/TWC.2014.2377121

the system. Indeed, the transient-performance of the systems was seldom addressed in the literature. Since the congestion of the RACHs occurs in a short interval, the performance metrics of the RACHs should be collected for the period of time between the activation of the first M2M device and the (successful or unsuccessful) completion of the last random access procedure triggered by the last M2M device [3]. The steady-state analysis cannot provide enough information to improve the operation of the RACHs. Therefore, it is worthwhile to conduct transient analysis of the RACHs under bursty arrivals.

In this paper, we present an analytical methodology to model and analyze the RACHs with bursty arrivals in OFDMA cellular networks. More specifically, this work provides three major contributions.

1) We provide new probabilistic "tools" for the analysis of multi-channel RACHs. Specifically, using non trivial combinatorics (involving the necessary extension of the Stirling Numbers of the Second Kind), we believe this paper is the first to derive an *exact* expression for the probability distribution of the number of successful transmission attempts over multiple available channels (Random Access Opportunities (RAO)). As such, our main theorem (Theorem II.3) can be extremely useful to networking and random access researchers, even beyond the specific application to OFDMA systems tackled in this work.
2) We derive a simple and computationally convenient *drift approximation* for the analysis of OFDMA RACHs in transient conditions and with non stationary arrivals; in doing this, we provide a more formal and systematic justification of the iterative methods used in our past works [13], [14].
3) We use the drift approximation to model a *realistic* OFDMA system based on the detailed LTE random-access procedure specified in the 3GPP standard [1]. We use the approximation model to estimate the OFDMA RACH performance with generally distributed bursty arrivals, including collision/success probability, average access delay, and Cumulative Distribution Function (CDF) of the number of preamble transmissions and access delay for the successfully accessed MSs.

The rest of the paper is organized as follows. Section II describes the analytical methodology adopted in this paper. The system model considered in this paper is defined in Section III. An analytical model and the performance metrics of the RACHs with bursty arrivals are presented. The numerical results are given in Section IV. Section V summarizes the conclusion.

A. Related Work

OFDMA is the major technology adopted by the next generation cellular networks such as 3GPP LTE-A or IEEE 802.16m. In OFDMA, time is divided into fixed-length radio frames. Each radio frame consists of multiple sub-frames. Preamble transmissions are restricted to specific sub-frames referred as random-access slots [15]. In OFDMA, each random-access slot contains multiple RAOs. Hence, the behavior of RACHs in an OFDMA wireless network is similar to a multi-channel slotted ALOHA system [8].

Several analytical models have been proposed to derive the throughput [4], [11], [12], the collision probability [5], [7], and the access delay [4], [5], [7], [9] of the next generation cellular networks. A comprehensive survey and comparison of the existing research proposals to solve the RACH congestion problem has been presented in [2]. The authors classify existing solutions into four categories: MAC layer optimizations; separate resource assignment for M2M and H2H traffic; distribute the arrivals along time resorting to random counters; and hybrid approaches combining all these concepts to improve the performance of the RACHs. In [5], the authors presented an analytical model to evaluate the access success probability and the access delay for UMTS system. Simulation results of access success probability for GSM network were provided in [6]. The access delay and the throughput of an OFDMA system was studied in [4], [7] via computer simulations. The RACH throughput and the access delay considering the periodic access characteristic of an OFDMA system is presented in [4]. The effect of backoff algorithms for RACHs in UMTS-LTE and IEEE 802.16 systems was investigated in [5], [7]. In [9], simulation results of the CDF of the access delay and the utilization of RACHs for IEEE 802.16 system were presented. In [2], the authors investigated the energy efficiency of the RACHs through computer simulations.

A significant amount of work has been done to investigate and improve the operation of the RACHs for next generation cellular networks. However, comparing different approaches is not straightforward, because researchers often highlight the advantages of their approach by considering different performance evaluation scenarios. To align all the efforts in the same direction, 3GPP [3] defines six key performance metrics and two application scenarios to evaluate the performance of the novel proposals being designed today [2]. The identified key performance metrics include i) (preamble) collision probability, ii) access success probability, iii) statistics of the number of preamble transmissions, iv) statistics of access delay, v) statistics of simultaneous preamble transmissions, and vi) statistics of simultaneous data transmissions. The collision probability is the ratio between the number of occurrences when two or more M2M devices send a random access attempt using exactly the same preamble and the overall number of opportunities (with or without access attempts) in the period. The access success probability is the probability that an MS successfully completes the random access procedure within the maximum number of preamble transmissions. The statistics of number of preamble transmissions is the CDF of the number of preamble transmissions to perform a random access procedure, for the successfully accessed M2M devices. The statistics of access delay is the CDF of the delay for each random access procedure between the first random access attempt and the completion of the random access procedure, for the successfully accessed M2M devices. The statistics of simultaneous preamble transmissions is the CDF of the number of M2M devices that transmit preamble simultaneously in an random-access slot [3]. The statistics of simultaneous data transmissions is the CDF of the number of M2M devices that transmit pilot or pilot and data simultaneously in an random-access slot. The statistics of simultaneous preamble transmissions and the statistics of simultaneous data transmissions are for UMTS FDD only. This work addresses four (i to iv) key performance metrics defined by 3GPP. In addition to the four performance metrics, we further derive the performance of average access delay for the successfully accessed MSs. Non-synchronized and synchronized application scenarios are considered. Note that non-synchronized application scenario is modeled by traffic model 1, in which M2M devices access the network uniformly over a period of time, whereas the synchronized application scenario (or, the event-driven M2M communication considered in [16]) is modeled by traffic model 2, in which a large amount of M2M devices access the network following a time-limited Beta distribution. The effects of power ramping and radio channels (e.g., path-loss, fading, inter-cell interference, etc.) are characterized by a time-dependent preamble detection probability [3].

The transient analysis of the slotted ALOHA protocol was first presented in [10]. The authors proposed a continuous-state diffusion process to approximate the number of backlogged users as a function of time and showed that the time-dependent throughput of the slotted ALOHA protocol is a Gauss-Markov process with time-dependent mean and variance [10]. However, it is not clear how to generalize the model to accommodate multi-channel slotted ALOHA systems with a time-varying retransmission probability due to backoff policy. The random access control of M2M communication systems with bursty traffic has been investigated in [14], [16]. In these studies, the burstiness nature of event-driven M2M traffic was modeled as batch arrivals in a single random-access slot in [14] or as a two-state interrupted Poisson process in [16]. In [14], an analytical model was presented to estimate the access success probability and the collision probability of RACHs in group paging observed during a target paging access interval. The performance metrics were derived based on the average number of contending and success devices estimated in each slot. In [16], a fast adaptive slotted ALOHA scheme was presented to adjust the transmission probability of a *p*-persistent slotted ALOHA system. Similar to the Pseudo Bayesian method presented in [17], both methods aimed to estimate the number of active devices in a slot. However, the main difference is that the number of active devices is estimated based on the access outcome in the previous random-access slot in [17] but on the access outcomes in past consecutive slots in [14]. For analytical tractability, the impact of backoff policy and the power-ramping effect of the random access procedure were neglected in both studies [14], [16].

II. Analytical Methodology

In OFDMA, time is divided into fixed-length radio frames. Each radio frame consists of multiple sub-frames. Preamble transmissions are restricted to specific sub-frames [15], which are referred as the random-access slots in the rest of this paper. In OFDMA, the random-access resource is determined in terms of RAOs. RAO is the random-access preamble allocated in each frequency bands in a random-access slot. As illustrated in Fig. 3, the total number of RAOs in a random-access slot is the number of random-access preambles allocated in each frequency band multiplied by the number of frequency bands in a random-access slot [3]. For simplicity, a single frequency band is considered herein. Hence, RAO and preamble are used interchangeably in the rest of the paper. An MS which generates a random-access attempt should wait for the next available random-access slot and transmit through a randomly chosen RAO.

The analytical methodology exploited in Section III to model the detailed LTE random access procedure [1] consists in approximating the stochastic behavior of the OFDM RACH operation with its *deterministic* "mean" trajectory, derived by solving the discrete-time *drift* vector equation associated to the stochastic process. The proposed approach is motivated by two facts. First, we aim to model transient (bursty) arrival processes, rather than steady-state conditions. Second, the number of successful transmissions in a slot, derived in Section II-B, results to be a non trivially distributed random variable $\mathcal{S}$, which makes a rigorous stochastic analysis cumbersome.

In the remainder of this section we describe (and provide the relevant theoretical foundation for) such a modeling approach, using a reference scenario of a *simplified* RACH operation which permits the reader to focus on the methodology without the need to enter, at least at this stage, in the supplementary technicalities of the LTE random access procedure (whose detailed analysis is postponed to Section III). This section is organized in three parts. In Section II-A we introduce and model the simplified RACH operation as a stochastic process. The stochastic model requires the knowledge of the distribution of the number of successful transmissions in a random-access slot, which is separately derived in Section II-B.[1] Finally, in Section II-C we show how such stochastic process can be approximated via a system of discrete-time drift equations, which can be easily (numerically) solved in transient conditions even for large scale systems.

A. Simplified RACH Operation: Stochastic Model

For the purpose of this section, let us consider a simplified OFDMA RACH operation: a more realistic setting devised to model the LTE random access procedure [1] will be introduced in the next Section III. Let us assume a regularly slotted RACH, where time slots are referred to with the index $i \geq 0$. For each random-access slot i, let $\mathcal{A}_i$ be a random variable which represents the number of newly arriving MS during random-access slot i. Let R be the number of RAOs made available by the BS in each random-access slot. Upon arrival, an MS will attempt to establish a connection with the BS by transmitting a preamble, i.e., by selecting a randomly chosen RAO among the R available RAOs. If the transmission is successful, the MS will not transmit anymore over the RACH. Conversely, if the transmission fails, the MS will try to retransmit in subsequent

[1] Indeed, this result appear non trivial, as it requires an extension of the Stirling Numbers of the second kind; as such, other than permitting the rigorous analysis of small scale systems, we believe that the distribution derived in this section may be useful even besides this specific paper, e.g., in other contexts where multiple resources (slots, channels, etc.) are randomly accessed by a pool of users.

slots, until a maximum number of transmissions $N_{PT\,\max}$ is reached. Moreover, let us further *neglect* for the purpose of this section the backoff process, and rather assume that a station which experiences a collision immediately retries a transmission in the subsequent slot.

Even in such overly simplified conditions, a rigorous model requires us to study the transient behavior of the multi-dimensional discrete-time stochastic process

$$\vec{\mathcal{M}_i} = \langle \mathcal{M}_i[1], \mathcal{M}_i[2], \cdots, \mathcal{M}_i[N_{PT\,\max}] \rangle, \tag{1}$$

where $\mathcal{M}_i[n]$, with $n \in (1, N_{PT\,\max})$, represents the number of MSs which are engaged in their n-th transmission at random-access slot i. Obviously, for any random-access slot i and retry index n, $\mathcal{M}_i[n]$ is a random variable. For $i \geq 0$, and for a general arrival process $\mathcal{A}_i$, the evolution of such process can be described by the stochastic recursion

$$\begin{cases} \mathcal{M}_{i+1}[1] & = \mathcal{A}_i \\ \mathcal{M}_{i+1}[2] & = \mathcal{M}_{i,F}[1] \\ \cdots & \quad \cdots \\ \mathcal{M}_{i+1}[n] & = \mathcal{M}_{i,F}[n-1] \\ \cdots & \quad \cdots \\ \mathcal{M}_{i+1}[N_{PT\,\max}] & = \mathcal{M}_{i,F}[N_{PT\,\max}-1] \end{cases} \tag{2}$$

where $\mathcal{M}_{i,F}[n]$ is the random variable which represents the number of MSs which experience a collision (their transmission fails) during their n-th transmission attempt in random-access slot i.

In order to determine the probability distribution of the r.v.s. $\mathcal{M}_{i,F}[n]$, we conveniently introduce a *complementary* random variable $\mathcal{M}_{i,S}[n]$, which represents the number of successful transmissions during random-access slot $i \geq 0$ by MSs engaged in their n-th transmission attempt, $n \in (1, N_{PT\,\max})$. Obviously, under the simplificative assumption that all MSs transmit in each random-access slot (i.e., no backoff), the sum of the number of successful and failed MS transmissions is equal to the number of MSs engaged in their n-th transmission, i.e.,

$$\mathcal{M}_{i,F}[n] + \mathcal{M}_{i,S}[n] = \mathcal{M}_i[n] \qquad \forall i \geq 0, n \in (1, N_{PT\,\max}). \tag{3}$$

Let us now assume that the state of the stochasting process $\vec{\mathcal{M}_i}$ at random-access slot i is given, and let us denote[2] this (deterministic) state as $\vec{M}_i = \langle M_i[1], \cdots, M_i[N_{PT\,\max}] \rangle$. Let us further denote with $M_i = \sum_{n=1}^{N_{PT\,\max}} M_i[n]$ the total number of MSs competing during random-access slot i in the considered state. Finally, let us now define with $S_R\ (M_i; k)$ the probability that exactly k out of the M_i transmitting MSs are successful, given that a total number of R RAOs are available. (This distribution will be derived in the next Section II-B). Then, for any given state $\vec{M}_i$, we can derive the joint distribution of the (correlated) random variables $\mathcal{M}_{i,S}[n]$ as:

$$P\left\{\mathcal{M}_{i,S}[1] = s_1, \cdots, \mathcal{M}_{i,S}[N_{PT\,\max}] = s_{N_{PT\,\max}} | \vec{\mathcal{M}_i} = \vec{M}_i\right\} = \sum_{k=0}^{M_i} S_R(M_i; k) \sum_{\substack{s1+s2+\cdots \\ +s_{N_{PT\,\max}}=k}} \frac{\binom{M_i[1]}{s_1} \cdots \binom{M_i[N_{PT\,\max}]}{s_{N_{PT\,\max}}}}{\binom{M_i}{k}}, \tag{4}$$

and hence, from (3), the complementary distribution of the r.v.s. $\mathcal{M}_{i,F}[n]$ involved in (2). Such r.v.s. $\mathcal{M}_{i,F}[n]$ depend only on the state of the process at time i. It readily follows that the stochastic process (1) which models the considered (simplified) OFDMA RACH operation is a Markov Chain, under the *supplementary* assumption that the arrival process $\mathcal{A}_i$ is either i) a sequence of independent (but not necessarily identically distributed, hence they can be time-dependent) random variables having general distribution, or ii) the arrival process $\mathcal{A}_i$ is a Markov chain itself. The first case is trivial. In the second case, it suffices to note that, owing to the specific structure of the stochastic process (1), the memory of the arrival process itself (namely, the number of arrivals occurred in the last slot) is "stored" in the state variable $M_i[1]$. Besides, note that the sum $\sum_{n=1}^{N_{PT\,\max}} \mathcal{M}_{i,S}[n]$ permits to compute the total number of successful transmissions at random-access slot i, i.e., MSs which succeed in connecting to the BS at random-access slot i and hence leave the RACH, whereas $\mathcal{M}_{i,F}[N_{PT\,\max}]$ yields the number of MSs which declare failure at slot i and will abort the connection attempt.

B. Probability Distribution of the Number of Successful Transmissions

In order to complete the stochastic model presented above, in this section we derive an explicit expression for the probability distribution $S_R\ (M_i; k)$ used in equation (4).

Let us focus on a given random-access slot, and, to simplify notation, let m be the number of MSs which select their transmission preamble among the R RAOs available in the random-access slot. Owing to collisions emerging when two or more MSs select a same RAO, the number of successful transmissions is in most generality a random variable $\mathcal{S}$ taking values in the range 0 to $\min(m; R)$. Let $S_R\ (m; k)$ be the probability distribution of $\mathcal{S}$, i.e. the probability that *exactly* k out of the m MSs successfully transmit in a random-access slot comprising R RAOs. The problem of deriving such a probability distribution can be cast into a *Balls & Bins* combinatorics problem. We have m "balls" (number of MSs which transmit preamble); each ball randomly chooses one among R available "bins" (RAOs), and we wish to find the probability that k bins contain exactly a single ball (successful transmissions). To solve this problem, we start from the following simple Lemma.

Lemma II.1: Consider $R \geq 1$ bins and $m \geq 1$ balls. Each ball is independently placed in a randomly chosen bin. Let $\mathcal{U}$ be the random variable representing the resulting number of non empty bins. Then $\mathcal{U}$ has the following probability distribution:

$$U_R(m; x) = \frac{\left\{ {m \atop x} \right\} \binom{R}{x} x!}{R^m}, \qquad 1 \leq x \leq \min(m, R), \tag{5}$$

where $\left\{ {m \atop x} \right\}$ are Stirling Numbers of the Second Kind. Moreover, the expected value of $\mathcal{U}$ is given by

$$E[\mathcal{U}] = R\left(1 - \left(1 - \frac{1}{R}\right)^m\right). \tag{6}$$

Proof: The lemma reduces to a counting exercise. We recall that $\left\{ {m \atop x} \right\}$ expresses the number of ways to partition a set of

[2] For notational convenience, we refer to random quantities with calligraphic notation $\mathcal{M}_i[n]$, whereas we refer to given (deterministic) quantities with plain notation $M_i[n]$.

m elements (labeled balls) into x non empty subsets (unlabeled bins). As well known, Stirling numbers of the second kind can be readily computed via the closed form expression

$$\left\{ m \atop x \right\} = \frac{1}{x!} \sum_{i=0}^{x-1} (-1)^i \binom{x}{i} (x-i)^m, \tag{7}$$

or via the convenient recursion

$$\left\{ m \atop x \right\} = \left\{ m-1 \atop x-1 \right\} + x \cdot \left\{ m-1 \atop x \right\} \tag{8}$$

using, as initial conditions, $\left\{ m \atop 1 \right\} = 1$ and $\left\{ m \atop m \right\} = 1$. $\binom{R}{x}$ yields the number of ways in which exactly x bins are chosen out of R total bins, and $x!$ is the number of ways we can label the chosen bins. In summary, the numerator in (5) provides the number of ways in which m labeled balls fall into exactly x labeled bins out of R available ones. The probability distribution is finally derived by dividing for the total number R^m of ways to distribute m labeled balls across R labeled bins. Finally, (6) can be computed via a direct argument, by exploiting the basic fact that, also for non independent random variables X_i (indeed our case below), $E[\sum_i X_i] = \sum_i E[X_i]$. Hence it suffices to describe a single bin via the random variable $X_i \in \{0,1\}$ which assumes value 1 when the bin is non empty; trivially note that $E[X_i] = 1-(1-1/R)^m$, and multiply by the total number R of bins to obtain the mean value $E[\mathcal{U}]$ in (6). □

In our setting, Lemma II.1 has the following interpretation: $U_R(m;x)$ is the probability that exactly x out of R RAOs are used by *one or more* transmissions generated by the m MSs. Unfortunately, it does not tell us how many of such x RAOs do envision *exactly* one transmission. This further step is not straightforward,[3] and requires to extend Stirling numbers as follows.

Lemma II.2: Let $\Psi^m_{x,j}$ be the number of ways we can partition $m \geq 1$ elements into $x \geq 1$ subsets, $j \in (0,x)$ of which contain *more than one* element. The following recursive expression for $\Psi^m_{x,j}$ holds:

$$\Psi^m_{x,j} = j\Psi^{m-1}_{x,j} + (x-j+1)\Psi^{m-1}_{x,j-1} + \Psi^{m-1}_{x-1,j}, \tag{9}$$

with initial conditions $\Psi^m_{m,0} = 1$, $\Psi^{m=1}_{1,0} = 1$, $\Psi^{m=1}_{1,1} = 0$, and $\Psi^{m>1}_{1,1} = 1$. Note that $\forall m \geq 1, \sum_{j=0}^{x} \Psi^m_{x,j} = \left\{ m \atop x \right\}$.

Proof: The recurrence (9) can be envisioned as a generalization of (8), with the fundamental difference that we now explicitly count and track the number of singletons in a set partition. Initial conditions are self-evident. A single element can only be partitioned in a single unitary set, hence $\Psi^1_{1,0} = 1$, whereas $\Psi^1_{1,1} = 0$. There is only one way to partition $m > 1$ elements in a single set, and the resulting set has size greater than 1, hence $\Psi^{m>1}_{1,1} = 1$. Moreover, there is only one way to partition m elements into m necessarily singleton sets, hence $\Psi^m_{m,0} = 1$. The recurrence (9) follows from an enumerative argument. There are at most three possible ways to partition m elements into x sets, of which j have two or more elements (multitons, in short):

1) partition $m-1$ elements into x sets, j of which multitons, and then add a last element to a multiton. This can be done in $j\Psi^{m-1}_{x,j}$ ways;
2) partition $m-1$ elements into x sets, $j-1$ of which multitons, and then add a last element to a singleton, thus increasing the number j of multitons of one unit. This can be done in $(x-(j-1))\Psi^{m-1}_{x,j-1}$ ways;
3) partition $m-1$ elements into $x-1$ sets, j of which multitons, and then add the last element as a new singleton, thus increasing the number x of sets of one unit, leaving the number j of multitons unvaried. This can be done in $\Psi^{m-1}_{x-1,j}$ ways.

Finally, the relation $\sum_{j=0}^{x} \Psi^m_{x,j} = \left\{ m \atop x \right\}$ with the Stirling numbers can be trivially proved via induction (straightforward proof omitted). The intuition behind such a relation is that the terms $\Psi^m_{x,j}$ do not change (of course!) the number of ways m elements can be partitioned into x sets, but further detail the ways in which the total number of possible partitions $\left\{ m \atop x \right\}$ further subdivide in terms of number of multitons j and complementary number of singletons $x-j$. □

With the extension of the Stirling number of the second kind above introduced, we are now able to give an exact answer to our initial question, namely what is the probability distribution $S_R(m;k)$ of the r.v. $\mathcal{S}$ which models the fact that *exactly* k out of the m MSs successfully transmit in a random-access slot comprising R RAOs.

Theorem II.3: The probability $S_R(m;k)$ that exactly k out of $m \geq 1$ MS successfully transmit in a random-access slot comprising R RAOs is given by:

$$\begin{aligned} S_R(m;k) &= \sum_{x=k}^{\min\left(\left\lfloor \frac{m+k}{2} \right\rfloor; R\right)} U_R(m;x) \frac{\Psi^m_{x,x-k}}{\left\{ m \atop x \right\}} \\ &= \sum_{x=k}^{\min\left(\left\lfloor \frac{m+k}{2} \right\rfloor; R\right)} \frac{\Psi^m_{x,x-k} \binom{R}{x} x!}{R^m}, \\ & \qquad 0 \leq k \leq \min(m;R). \end{aligned} \tag{10}$$

Moreover, the expected number $E[\mathcal{S}]$ of successful transmissions in a random-access slot is given by:

$$E[\mathcal{S}] = m \left(1 - \frac{1}{R}\right)^{m-1}. \tag{11}$$

Proof: The first part of the theorem is a straightforward consequence of Lemmas II.1 and II.2. We simply apply the total probability theorem by conditioning on the number x of RAOs which envision a transmission, and by computing the probability $\Psi^m_{x,x-k}/\left\{ m \atop x \right\}$ that exactly k of such x transmissions are successful. The lower range of the sum is obviously $x = k$ as at least k RAOs must be used by the k successful transmissions. The upper range of the sum comes from the remark that the number m of MSs must be greater or equal than the number x of

[3] For a concrete simple example of the underlying subtleties, consider the case of 4 balls, $\{a,b,c,d\}$ in two bins. There are $\left\{ 4 \atop 2 \right\} = 7$ possible cases. Four cases have a singleton set, namely $\{\{a\},\{b,c,d\}\}$, $\{\{b\},\{a,c,d\}\}$, $\{\{c\},\{a,b,d\}\}$, $\{\{d\},\{a,b,c\}\}$. The other three cases do not have any singleton: $\{\{a,b\},\{c,d\}\}$,$\{\{a,c\},\{b,d\}\}$, $\{\{a,d\},\{b,c\}\}$. The apparently "obvious" conditioning approach consisting in "removing" the two head of the line balls, and "redrawing" the remaining two would thus grossly fail (would yield equal probability 1/2 instead of 3/7 and 4/7).

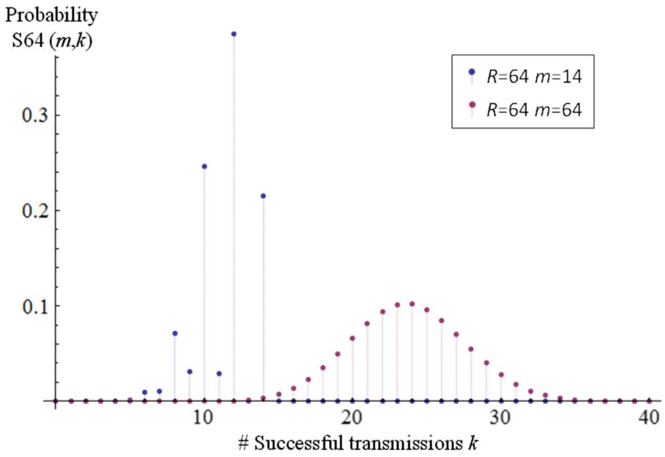

Fig. 1. Distribution of the number of successful transmissions in a given slot, equation (10), $R = 64$, $m = 14, 64$.

used RAOs plus the number $x - k$ of RAOs which envision two or more transmissions, i.e. $m \geq x + (x-k) \rightarrow x \leq (m+k)/2$, with the floor operation motivated by the fact that x must be an integer value. More interesting is the derivation of the mean value, which again can be derived via a direct argument. Let us focus on a single RAO i, and let Y_i be a random variable which assumes value 1 if the RAO is used by *exactly* one transmission, and 0 otherwise. It readily follows that, for a single RAO,

$$E[Y_i] = \binom{m}{1}\frac{1}{R}\left(1-\frac{1}{R}\right)^{m-1} = \frac{m}{R}\left(1-\frac{1}{R}\right)^{m-1}. \qquad (12)$$

The expression (11) follows since $E[\mathcal{S}] = R \cdot E[Y_i]$, which holds also for non independent r.v.s. Y_i as it indeed is our case. □

Fig. 1 plots the probability distribution $S_R\ (m;k)$ of the number of successful transmissions in a random-access slot comprising a relatively large number of RAOs, specifically $R = 64$, for the two cases of $m = 14$ and $m = 64$ competing MSs. As intuitively expected, the distribution in the case of $m = 14$ is irregular. Indeed, since a collision must involve at least 2 MSs, we cannot have 13 successful transmissions; similarly, an odd number (11, 9, etc.) of successful transmissions is less likely to occur, as it would require 3 MSs to collide for a same RAO. Despite the irregularity, we see that the probability mass is concentrated around the mean value 11.4. Similar concentration around the mean value 23.7, but with a more regular Gaussian-like shape, occurs for the $m = 64$ case. In essence, Fig. 1 intuitively suggests that, for a relatively large R, the Probability Density Function (pdf) of the number of successful transmissions is practically concentrated around its mean value, for either m small and large. To formally quantify such concentration, we can further compute the *coefficient of variation* defined as $C_v = \sqrt{Var[\mathcal{S}]}/E[\mathcal{S}]$. For $R = 64$ and $m = 14$, $C_v = 0.168$; for $R = 64$ and $m = 64$, $C_v = 0.163$. This fact will turn out to be useful, since, as discussed next, a rigorous queueing analysis in transient conditions appears computationally cumbersome.

C. Drift Approximation

In principle, the probability distribution of the number of successful transmissions in a random-access slot, derived in the previous Section II-B, permits us to complete the stochastic model introduced in Section II-A. Indeed, the ability to compute $S_R\ (M_i;k)$ permits us to determine the conditional distribution (4), and, via (3), further derive the transition probabilities for the discrete-time Markov Chain (2). Thus, we would be able to study the system evolution using standard Markovian techniques. Nevertheless, the solution (in transient conditions) of such system would be non trivial and burdened by significant computational complexity due to state space explosion even for a moderate value of $N_{PT\,\max}$, as well as lack of a closed form expression for (4), and an exact analysis would be cumbersome even in an overly simplified scenario such as the one considered so far.

To circumvent such issues, and motivated by the fact that practical values of R are relatively large (several tens) and in such conditions the distribution (10) is consistently concentrated around its mean value (Fig. 1), in this paper we propose to approximate the above described discrete time stochastic process $\vec{\mathcal{M}}_i$ with a *deterministic* discrete time system $\vec{M}_i$ which describes the transient evolution of the *drift* of the stochastic process itself. This approach is frequently exploited in the literature, see for instance the surveys [18], [19] and is shown to asymptotically converge to *exact* results for finite state space systems under mild assumptions (Lipschitz conditions). Indeed, our own results will show a very accurate matching between simulation and analytical results even for a non particularly large value of R (we used $R = 54$ in conformance to the LTE specification [1]), and for the much more elaborated and realistic system model presented in the next Section III.

The proposed approximation is established as follow. As discussed in Section II-A, in the assumption of independent (or Markovian) arrivals $\mathcal{A}_i$, the stochastic process $\vec{\mathcal{M}}_i$ introduced in (1) is a discrete-time Markov Chain. The *drift* $\vec{d}_i$ of this Markov chain at time i, is defined, for every given state $\vec{M}_i$, as the conditional expectation

$$\vec{d}_i(\vec{M}_i) = E[\vec{\mathcal{M}}_{i+1} - \vec{\mathcal{M}}_i | \vec{\mathcal{M}}_i = \vec{M}_i]. \qquad (13)$$

Our approximation consists in modeling the evolution of the process by means of the deterministic, "mean," trajectory specified by the discrete time drift equation

$$\vec{M}_{i+1} = \vec{d}_i(\vec{M}_i) + \vec{M}_i = E[\vec{\mathcal{M}}_{i+1} | \vec{M}_i], \qquad (14)$$

where the rightmost expression immediately follows from (13). Owing to (2), we can rewrite (14) in expanded form, as a system of $N_{PT\,\max}$ equations:

$$M_{i+1}[n] = \begin{cases} E[\mathcal{A}_i] & n = 1; \\ E\left[\mathcal{M}_{i,F}[n-1] | \vec{M}_i\right] & 2 \leq n \leq N_{PT\,\max}. \end{cases} \qquad (15)$$

For compact notation, let us now define the deterministic "correspondents" of the random variables representing the successful and failed MS transmissions per each random-access slot i and per retry attempt n:

$$M_{i,F}[n] = E\left[\mathcal{M}_{i,F}[n]|\vec{M}_i\right];$$
$$M_{i,S}[n] = E\left[\mathcal{M}_{i,S}[n]|\vec{M}_i\right]. \tag{16}$$

Owing to (4), and noting that the expected number of successful nth preamble transmissions at the ith random-access slot depends only on the number $M_i[n]$ of MS, and the total number $M_i = \sum_{n=1}^{N_{PT\,\max}} M_i[n]$ of MS, $M_{i,S}[n]$ is readily computed as:

$$M_{i,S}[n] = \frac{M_i[n]}{M_i} E[\mathcal{S}|M_i] = M_i[n]\left(1-\frac{1}{R}\right)^{M_i-1} \tag{17}$$

where $E[\mathcal{S}|M_i]$ is computed using (11). This expression can be further well approximated[4] by

$$M_{i,S}[n] \approx M_i[n] e^{-\frac{M_i}{R}}. \tag{18}$$

Moreover, since (3) yields the obvious equality $M_{i,F}[n] = M_i[n] - M_{i,S}[n]$, we can express $M_{i,F}[n]$ as:

$$\begin{aligned} M_{i,F}[n] =& M_i[n]\left(1-\left(1-\frac{1}{R}\right)^{M_i-1}\right) \\ \approx& M_i[n]\left(1-e^{-\frac{M_i}{R}}\right). \end{aligned} \tag{19}$$

By substituting (19) in (15), we obtain the final system of drift equations

$$M_{i+1}[n] = \begin{cases} E[\mathcal{A}_i] & n=1 \\ M_i[n-1]\left(1-e^{-\frac{M_i}{R}}\right) & 2 \le n \le N_{PT\,\max}. \end{cases} \tag{20}$$

Given an initial state, and a given arrival pattern, the recursion (20) can be numerically computed in a very efficient manner.[5] Note that, in most generality, the resulting solution, namely the time-varying vector $\vec{M}_i$, is real-valued, and identifies an approximating (discrete-time) deterministic trajectory for the original process (1) whose stochastic evolution is described by (2).

III. System Model

We are now ready to tackle the model of a more realistic OFDMA-based scenario. Our goal is to extend the above sketched analytical methodology to deal with the real-world implementation constraints of the random-access procedure. For concreteness, the LTE random-access procedure specified in 3GPP standard [1] is considered in the rest of the paper. The LTE random-access procedure can be divided into the steps of *preamble transmission*, *message transmission*, and *random backoff*, which are elaborated as below [1].

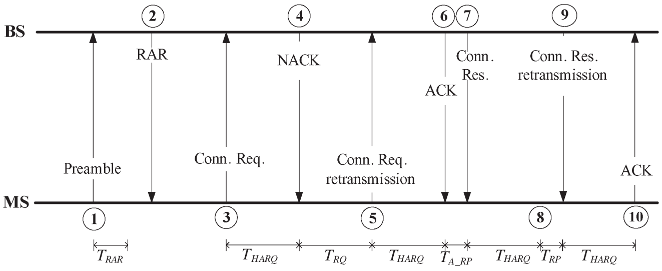

Fig. 2. Random access flow exchanges between the BS and MS.

A. Preamble Transmission

Initially, the MS transmits a preamble through a common channel shared by all MSs ((**1**) in Fig. 2). The BS correlates the received signal with the set of preamble sequences reserved for the cell and transmits a response message indicating the detected preamble(s) ((**2**) in Fig. 2). Each response message carries the backoff window size (unit: sub-frame) and one or more random access responses (RARs). Each RAR carries the identity (ID) of the preamble selected by the MS, the information to be used by the MS to adjust the uplink timing, and a dedicated uplink resource reserved for the MS to perform message transmission. Note that the maximal number of RARs that can be carried in a random-access response window, N_{UL}, is limited [1]. The MSs who successfully finish the step of *preamble transmission* have their preamble transmission not collided, detected by the BS, and indicated in RARs.

Considering the preamble detection probability, p_n, the total number of MSs whose n-th preamble transmission is detected by the BS is refined as $M_i[n]e^{-\frac{M_i}{R}}p_n$. Among these detected MSs, up to N_{UL} MSs can be acknowledged. Hence, $M_{i,S}[n]$ in Eq. (19) can be refined as

$$M_{i,S}[n] = \begin{cases} M_i[n]e^{-\frac{M_i}{R}}p_n, & \text{if } \sum_{n=1}^{N_{PT\,\max}} M_i[n]e^{-\frac{M_i}{R}}p_n \le N_{UL}, \\ \frac{M_i[n]e^{-\frac{M_i}{R}}p_n}{\sum_{n=1}^{N_{PT\,\max}} M_i[n]e^{-\frac{M_i}{R}}p_n} N_{UL}, & \text{otherwise.} \end{cases} \tag{21}$$

B. Message Transmission

The success MSs will receive the RAR message, and will transmit a *Connection Request* message carrying the ID through the dedicated uplink resource to the BS ((**3**) in Fig. 2). Non-adaptive hybrid automatic repeat request (HARQ) is subsequently enabled to protect the signaling exchange of the message transmission. In response, the BS transmits an HARQ acknowledgment (ACK) ((**6**) in Fig. 2) or negative-acknowledgment (NACK) ((**4**) in Fig. 2) after T_{HARQ} sub-frames. The BS waits for T_{A_RP} sub-frames and transmits a *Connection Response* message after an ACK indicating that *Connection Request* message is successfully received ((**7**) in

[4] The approximation readily follows by noting that, for x small, $e^{-x} \approx 1-x$ (first term of the Taylor Expansion around point 0). Thus, for large R, $e^{-M_i/R} = (e^{-1/R})^{M_i} \approx (1-1/R)^{M_i} \approx (1-1/R)^{M_i-1}$ (the last step holds as, large R, $(1-1/R)^{-1} \approx 1$).

[5] Of course, if the system operation has different solutions which depend on different initial conditions (multi-stability), the model will permit to derive only one of such solutions at a time; hence if the goal is to derive all the possible solutions, the model should be applied for different initial conditions.

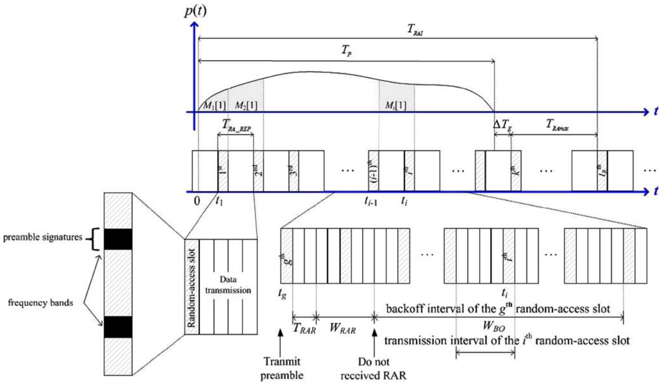

Fig. 3. The timing diagram of physical random-access transmission.

Fig. 2). The MS has to wait for T_{RQ} sub-frames and re-transmits a *Connection Request* message if it receives a NACK ((**5**) in Fig. 2). Similarly, the MS waits for T_{HARQ} sub-frames and transmits an ACK to the BS if it receives a *Connection Response* message ((**10**) in Fig. 2). The BS waits for T_{RP} sub-frames and re-transmits a *Connection Response* message ((**9**) in Fig. 2) if it didn't receive an ACK for the *Connection Response* message ((**8**) in Fig. 2). The HARQ retransmission of preamble and message can be up to N_{HARQ} times.

Let $p_{e,\mathrm{MSG}}$ be the error probability of the message transmission. The message transmission is failed if the transmission of *Connection Request* message exceeds N_{HARQ} times; or the *Connection Request* message is successfully transmitted but the transmission of *Connection Response* message exceeds N_{HARQ} times. Therefore, $p_{e,\mathrm{MSG}}$ is given by

$$P_{e,\mathrm{MSG}} = p_f^{N_{HARQ}} + \sum_{j=0}^{N_{HARQ}-1} p_f^j (1-p_f) p_f^{N_{HARQ}}. \qquad (22)$$

□

In what follows, we consider M MSs accessing a BS in an OFDMA-based wireless network during a short period of T_P sub-frames, and we investigate the transient-performance of RACHs observed over a random-access interval of T_{RAI} sub-frames. The arrival process of each of the M MSs follows an arbitrary pdf $p(t)$, as illustrated in the upper part of Fig. 3, where

$$\int_0^{T_P} p(t)\,dt = 1. \qquad (23)$$

The timing diagram of the physical random-access transmission of an OFDMA wireless network is illustrated in the lower part of Fig. 3. The starting point of the time axis $(t = 0)$ is chosen to be the arrival time of the first MS. Let t_i (unit: sub-frame) be the time at the beginning of the i-th random-access slot, as illustrated in the upper part of Fig. 3. t_i is given by

$$t_i = t_1 + (i-1)T_{RA\_REP}, \qquad (24)$$

where T_{RA_REP} is the interval between two successive random-access slots. The parameters of the LTE random-access procedure considered in this paper are summarized in Table I [3].

TABLE I
BASIC PARAMETERS USED IN THIS PAPER

Notation	Meaning	Numerical values
M	Total number of MS arriving during time interval T_P	5000, 10000, 30000
T_P	Arrival Period (sub-frames)	10000, 60000
W_{BO}	Backoff window size (sub-frames)	21
N_{UL}	Maximum number of MSs acknowledged within a random-access response window	$N_{RAR} \times W_{RAR}$
T_{RA_REP}	Interval between two successive random-access slots (sub-frames)	5
R	Total number of preambles in a random-access slot	54
N_{PTmax}	Maximum number of preamble transmission	10
N_{RAR}	Maximum number of RAR that can be carried in a response message	3
T_{CR}	Contention resolution timer (sub-frames)	48
W_{RAR}	Length of the random-access response window (sub-frames)	5
p_f	HARQ retransmission probability for a *Connection Request* message and a *Connection Response* message	10%
N_{HARQ}	Maximum number of HARQ transmissions for *Connection Request* and *Connection Response*	5
p_n	Preamble detection probability of the n-th preamble transmission	$p_n = 1 - \frac{1}{e^n}$
T_{RAR}	Processing time required by the BS to detect the transmitted preambles (sub-frames)	2
T_{HARQ}	Time interval required for receiving HARQ ACK (sub-frames)	4
T_{RQ}	Gap of *Connection Request* message retransmission (sub-frames)	4
T_{A_RP}	Gap of Monitor *Connection Response* message (sub-frames)	1
T_{RP}	Gap of *Connection Response* message retransmission (sub-frames)	1

Let N_{RAR} be the maximum number of RAR that can be carried in a response message; W_{RAR} be the size of the random-access response window (unit: sub-frame); N_{UL} be the maximum number of MSs that can be acknowledged within the random-access response window $(N_{UL} = N_{RAR} \times W_{RAR})$; T_{RAR} be processing delay required by the BS to transmit the RAR; $N_{PT\,\max}$ be the maximum number of preamble transmissions for each MS; p_f be HARQ retransmission probability for a *Connection Request* message and a *Connection Response* message; and p_n be the detection probability of the nth preamble $(1 \le n \le N_{PT\,\max})$.

C. Random Backoff

The average number of MSs involved in their first transmission is given by the average number of activate MSs

$$M_i[1] = M \int_{t_{i-1}+1}^{t_i+1} p(t)\,dt. \qquad (25)$$

Conversely, the MS for which the preamble or the message transmission fails, will perform a random backoff before retransmitting. The number of contending MSs that transmit nth ($n > 1$) preamble at the ith ($i > 0$) random-access slot, $M_i[n]$, contains two parts. The first part originates from the MSs in which the $(n-1)$th preamble transmissions at the gth random-access slot are failed (i.e., $M_{g,F}[n-1]$). Among these failed MSs, $\alpha_{g,i}$ of them perform random backoff and transmit the nth preambles at the ith random-access slot. Since these failed MSs was uniform backoff within backoff window, W_{BO}, g has multiple possibility ($G_{\min} \leq g \leq G_{\max}$).The second part originates from the MSs that transmit the $(n-1)$th preamble at the jth random-access slot ($J_{\min} \leq j \leq J_{\max} < i$); finish the preamble transmission; and the message transmission is failed (i.e., $M_{j,S}[n-1]p_{e,\text{MSG}}$). Among these failed MSs, $\beta_{j,i}$ of them will transmit the nth preambles at the ith random-access slot. Therefore, $M_i[n]$, for $2 \leq n \leq N_{PT\,\max}$, is given by

$$\begin{aligned} M_i[n] &= \sum\nolimits_{g=G_{\min}}^{G_{\max}} \alpha_{g,i} M_{g,F}[n-1] \\ &\quad + \sum\nolimits_{j=J_{\min}}^{J_{\max}} \beta_{j,i} p_{e,MSG} M_{j,S}[n-1] \\ &\approx \sum\nolimits_{g=G_{\min}}^{G_{\max}} \alpha_{g,i} M_{g,F}[n-1]. \end{aligned} \tag{26}$$

In general, $p_{e,MSG}$ is quite small and $\beta_{j,i}$ is less than 1. Hence, the second term in (26) for $2 \leq n \leq N_{PT\,\max}$ can be neglected.

$\alpha_{g,i}$, $G_{\min}$, and $G_{\max}$ in Eq. (26) are unknown parameters and can be derived based on the timing diagram given in Fig. 3 (see Appendix A).

D. Performance Metrics of the RACH

In this paper, the collision probability, access success probability, average access delay, CDF of number of preamble transmissions, and CDF of access delay, as suggested by 3GPP TR 37.868 [3], are chosen as the performance metrics to evaluate the performance of RACHs with bursty arrivals in OFDMA wireless networks. The performance metrics are derived within random-access interval reserved for all of M MSs to transmit their preambles T_{RAI}.

Before going into details, we first represent T_{RAI} in terms of the number of random-access slots, I_R. I_R is given by

$$I_R = \left\lfloor \frac{T_{\text{RAI}}}{T_{\text{RA\_REP}}} \right\rfloor. \tag{27}$$

As shown in Fig. 3, T_{RAI} is equal to the arrival time of the last arrived MS (T_P) plus the time used by the MS to wait for the next available random-access slot (ΔT_E) (unit: subframe) and the maximum period of time required by the MS to complete its random-access procedure ($T_{RA\,\max}$). That is,

$$T_{RAI} = T_P + \Delta T_E + T_{RA\,\max}. \tag{28}$$

From Fig. 3, it can be found that $T_P + \Delta T_E - t_1$ is a multiple of T_{RA_REP}. Hence, we can have

$$\Delta T_E = \begin{cases} 0, & \text{if } [(T_P - t_1) \bmod T_{RA\_REP} = 0], \\ T_{RA\_REP} - [(T_P - t_1) \bmod T_{RA\_REP}], & \text{otherwise,} \end{cases} \tag{29}$$

where $[x \bmod y]$ denotes the remainder of x divided by y. In each preamble transmission, the MS may spend up to $(T_{RAR} + W_{RAR} + W_{BO})$ sub-frames. Hence, $T_{RA\,\max}$ is given by

$$\begin{aligned} T_{RA\,\max} &= 1 + (N_{PT\,\max} - 1) \\ &\quad \cdot \left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rceil T_{RA\_REP}. \end{aligned} \tag{30}$$

1) Collision Probability: The collision probability, P_c, is defined as the ratio between the number of occurrences when two or more MSs send the same preamble with same frequency band and the overall number of RAOs (with or without access attempts) in the period [3]. That is, P_c is the ratio between the number of collided RAOs and the overall number of RAOs. The number of collided RAOs is equal to the total number of RAOs minus the number of success and idle RAOs [14]. Hence, P_c is given by

$$P_C = \frac{\sum\limits_{i=1}^{I_R} \left(R - M_i e^{-\frac{M_i}{R}} - R e^{-\frac{M_i}{R}} \right)}{I_R R}. \tag{31}$$

2) Access Success Probability: The access success probability, P_S, is the probability that an MS successfully completes the random-access procedure within the maximal number of preamble transmissions [3]. That is, P_S is the ratio between total number of successfully accessed MSs and the total number of MSs arrived in T_P. The number of successfully accessed MSs that transmit nth preamble at the ith random-access slot is equal to $M_{i,S}[n](1 - p_{e,\text{MSG}})$, where $p_{e,\text{MSG}}$ is the fail probability of the message transmission given in (22). Hence, P_S is given by

$$P_S = \frac{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n](1 - p_{e,\text{MSG}})}{M} \approx \frac{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n]}{M}. \tag{32}$$

3) Average Access Delay for the Successfully Accessed MSs: The average access delay for the successfully accessed MSs (unit: sub-frame), $\overline{D_a}$, is the ratio between the total access delay for all of the successfully accessed MSs and the total number of the successfully accessed MSs. $\overline{D_a}$ is given by

$$\overline{D_a} = \frac{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n](1 - p_{e,\text{MSG}})\overline{T_n}}{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n](1 - p_{e,\text{MSG}})} = \frac{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n]\overline{T_n}}{\sum\limits_{i=1}^{I_R} \sum\limits_{n=1}^{N_{PT\,\max}} M_{i,S}[n]}, \tag{33}$$

where $\overline{T_n}$ is the average access delay of a successfully accessed MS that transmits exactly n preambles. $\overline{T_n}$ contains the time required by the MS to transmit the first preamble, re-transmit the $(n-1)$ preamble(s), wait for the processing of the BS, and finish the message transmission (see Appendix B).

4) CDF of the Number of Preamble Transmissions: Let $F(m)$ be the CDF of the number of preamble transmissions to perform a random-access procedure for the successfully accessed MSs [3], where m is the number of preamble transmissions. $F(m)$ is the ratio between the number of successfully accessed MSs which transmit no more than m preambles and the number of all successfully accessed MSs. It is given by

$$F(m) = \frac{\sum_{i=1}^{I_R}\sum_{k=1}^{m} M_{i,S}[k](1-p_{e,\mathrm{MSG}})}{\sum_{i=1}^{I_R}\sum_{k=1}^{N_{PT\max}} M_{i,S}[k](1-p_{e,\mathrm{MSG}})} = \frac{\sum_{i=1}^{I_R}\sum_{k=1}^{m} M_{i,S}[k]}{\sum_{i=1}^{I_R}\sum_{k=1}^{N_{PT\max}} M_{i,S}[k]}. \tag{34}$$

5) CDF of the Access Delay: The CDF of the access delay, $G(d)$, is the ratio between the number of the successfully accessed MSs whose access delay is not greater than d and the total number of successfully accessed MSs. Note that d starts from the first random-access attempt and ends at the completion of the random-access procedure. $G(d)$ is estimated by

$$G(d) = \frac{\sum_{i=1}^{I_R}\sum_{k=1}^{m_{\max}(d)} M_{i,S}[k](1-p_{e,\mathrm{MSG}})}{\sum_{i=1}^{I_R}\sum_{k=1}^{N_{PT\max}} M_{i,S}[k](1-p_{e,\mathrm{MSG}})} = \frac{\sum_{i=1}^{I_R}\sum_{k=1}^{m_{\max}(d)} M_{i,S}[k]}{\sum_{i=1}^{I_R}\sum_{k=1}^{N_{PT\max}} M_{i,S}[k]}. \tag{35}$$

where $m_{\max}(d)(m_{\max}(d) \in N)$ is the maximal number of preambles transmitted by an MS and it can be estimated by setting $n = m_{\max}(d)$ in Eq. (39) and let Eq. (39) $= d$. That is, $m_{\max} = \lfloor (d - T_{RAR} - W_{RAR} - \overline{T_{MSG}} - 1)/\overline{T_W} \rfloor + 1$.

IV. Numerical Results

Computer simulations were conducted on top of a C-based platform to verify the effectiveness of the proposed analytical model. The analytical results of collision probability P_C, success probability P_S, average access delay $\overline{D_a}$, CDF of the number of preamble transmissions $F(m)$, and CDF of access delay $G(d)$ were obtained from Eqs. (31)–(35), respectively. Symbols and lines were used to denote simulation and analytic results, respectively. In the simulations, each point represented the average value of 10^5 samples. Each sample was obtained by observing the outcomes of the OFDMA RACHs with bursty arrival for T_{RAI} random-access slots. Two traffic models defined in 3GPP TR37.868 [3] were adopted to verify the accuracy of the analytical formula derived in Section III. The system parameters used in the simulation were set based on 3GPP LTE [3] and their values are summarized in Table I.

Fig. 4 showed the average number of arrived MSs ($M_i[1]$), success MSs ($M_{i,S}$), and failed MSs ($M_{i,F}$) in ith random-access slot for traffic model 2 with $M = 5000$. The results of traffic model 1 can be found in [13]. For comparison, only selected value of i of the simulation results were shown on the figure. Fig. 4 is a light-loaded scenario since we reserved $R = 54$ in each random-access slot but the average number of arrived MSs is less than 6. In this scenario, the number of failed MSs is mainly resulted from the non-perfect detection probability p_n.

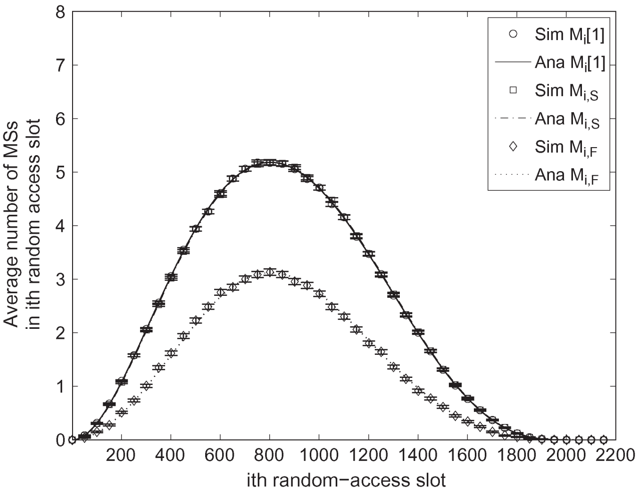

Fig. 4. Average number of arrived, success, and failed MSs in the ith random-access slot for traffic model 2 with 95% confidence interval, $M = 5000$.

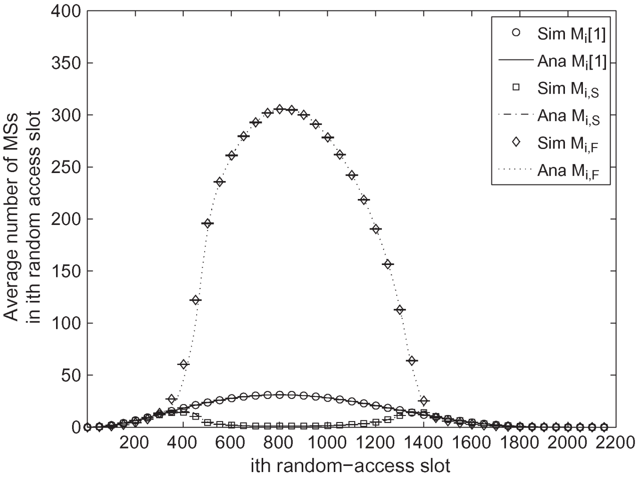

Fig. 5. Average number of arrived, success, and failed MSs in the ith random-access slot for traffic model 2 with 95% confidence interval, $M = 30\,000$.

Fig. 5 showed the average number of arrived, success, and failed MSs in ith random-access slot for traffic model 2 with $M = 30\,000$. The average number of failed MSs was significantly increased in such an overloaded scenario. The average number of success MSs was first increased thanks to the random backoff algorithm. It reached a local maximum value at $i = 380$ but then significantly reduced to zero during $i = 550$ to 1050 because of the excessive collisions resulted from the accumulated failed MSs. It started increasing again since $i = 1050$ because more MSs declared the random access failure and stopped contending for the RACHs. It reached another local maximum at $i > 1380$ and then decreased because only a few MSs were contending for the RACHs. It can be found in Figs. 4 and 5 that the proposed analytical model can accurately estimate the number of success MSs and failed MSs under different scenarios.

Figs. 6–8 showed the collision probability P_C, success probability P_S, and average access delay $\overline{D_a}$ of the two traffic models. It was found that the proposed analytical model can be applied for different kinds of traffic model. As the observation of $M_i[1]$, $M_{i,S}$, and $M_{i,F}$, the performance of the RACHs with traffic

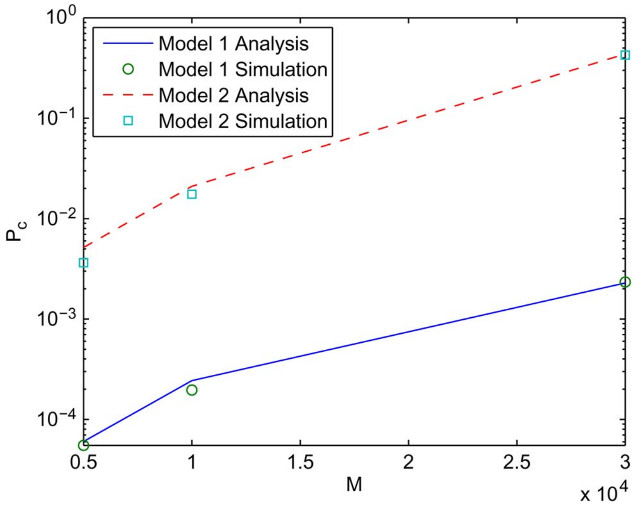

Fig. 6. Collision probability.

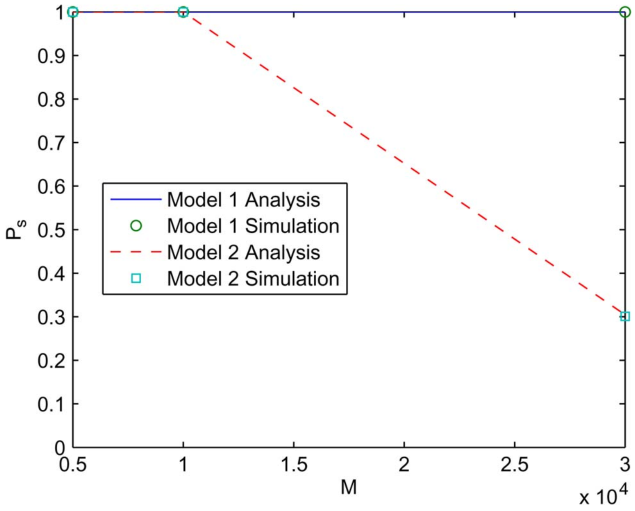

Fig. 7. Access success probability.

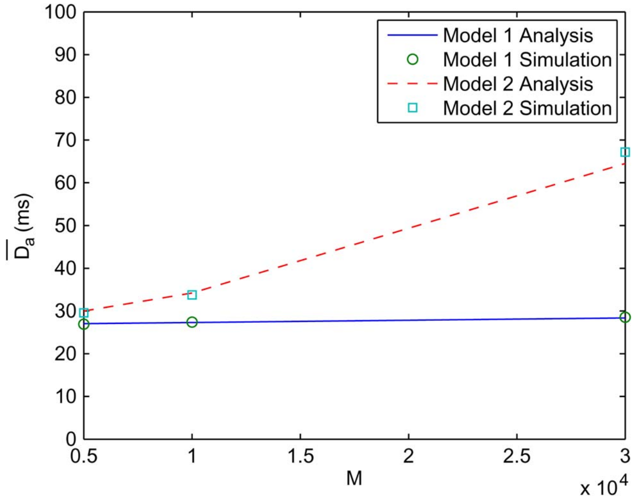

Fig. 8. Average access delay.

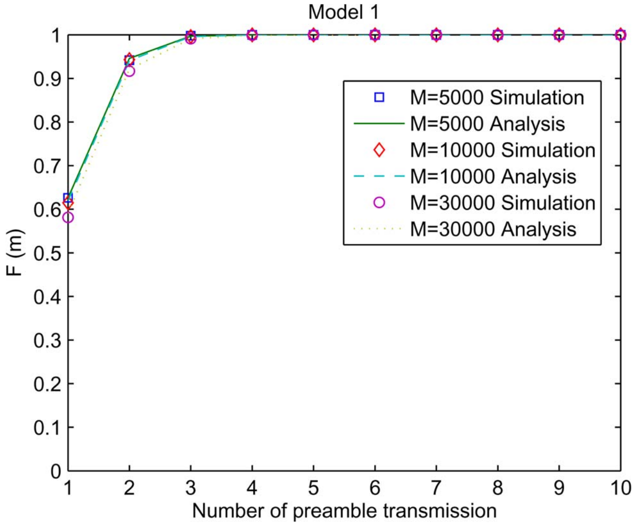

Fig. 9. CDF of preamble transmission of traffic model 1.

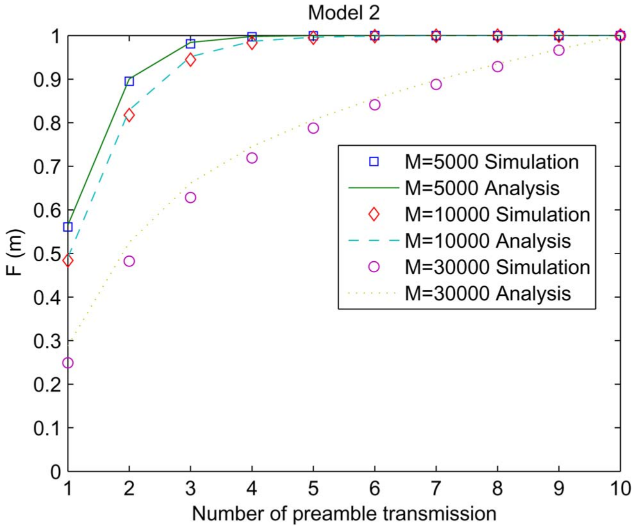

Fig. 10. CDF of preamble transmission of traffic model 2.

model 2 is worse than that with traffic model 1. It was also found that the network can accommodate up to 30 000 MSs in traffic model 1 but can only support 10 000 MSs in order to ensure 100% access success probability.

Figs. 9 and 10 showed the CDF of the number of preamble transmissions for the successfully accessed MSs, $F(m)$, for traffic models 1 and 2, respectively. It was found that the number of preamble transmissions in traffic model 2 is much higher than that in traffic model 1. In traffic model 1, 60% of the MSs can complete their random accesses by transmitting only one preamble and all MSs can complete their random accesses by transmitting up to four preambles.

Figs. 11 and 12 showed the CDF of the access delay for the successfully accessed MSs, $G(d)$, for traffic models 1 and 2, respectively. In two figures, the minimal access delay for an MS to complete its random accesses is 17 ms, which is the time required to transmit a preamble and a *Connection Request* message and to receive a *Connection Response* message. It was found that the analytical results of $G(d)$ were increased every $\overline{T_W}$ ms because $m_{\max}$ is a floor function of $\overline{T_W}$. Noticeable estimation errors between simulation and analytical results

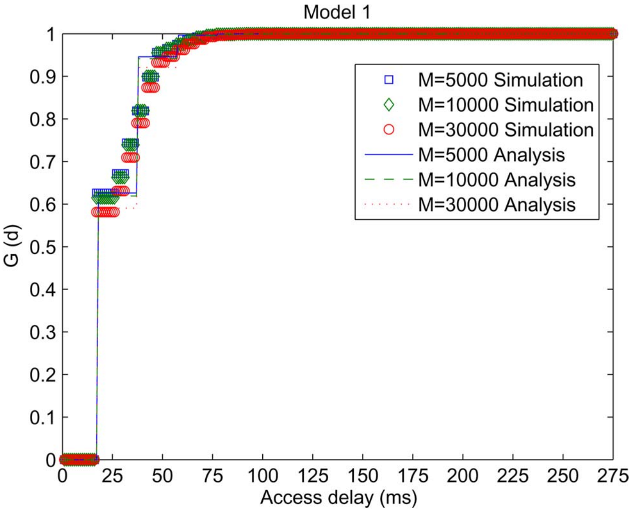

Fig. 11. CDF of access delay of traffic model 1.

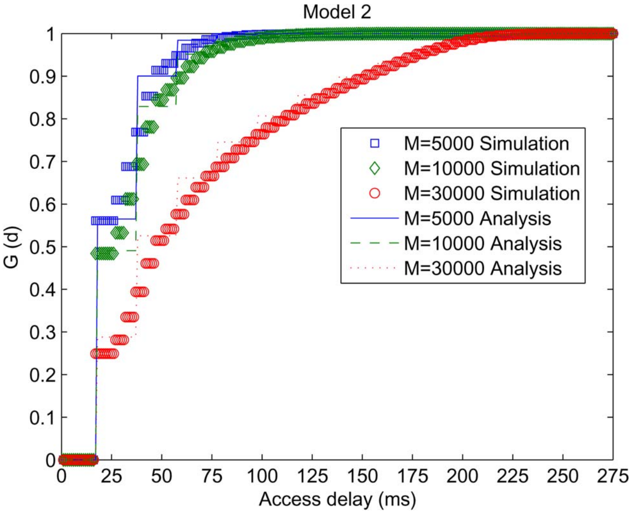

Fig. 12. CDF of access delay of traffic model 2.

were found in some regions of t because we used the average value of the access delay $\overline{T_n}$ to estimate $m_{\max}$ in deriving $G(d)$.

V. Conclusion

This paper presents an analytical model to investigate the transient behavior of the RACHs for an OFDMA system injected with bursty arrivals. The model suggests a way to estimate the number of contending MSs in each random-access slot based on the approximation formulas provided in [14] by considering the implementation constraints of the random-access procedure. The performance metrics of collision probability, success probability, average access delay, and the CDF of the number of preamble transmissions and access delay for the successfully accessed MSs are then derived. Numerical results showed that the proposed model can be applied for bursty arrivals with a general distribution. To the best of our knowledge, it is the first model that can accurately estimate the impact of bursty arrivals to the performance of the RACHs in OFDMA wireless networks.

Appendix A
Calculation of $\alpha_{g,i}$, $G_{\min}$, and $G_{\max}$

As shown in Fig. 3, the MSs which transmit their $(n-1)$th preambles at the gth random-access slot will recognize their random-access failure after T_{RAR} plus W_{RAR} subframes. The failed MSs perform random backoff and transmit the nth preamble when the backoff counter decreases to zero. Let W_{BO} be the backoff window size (unit: sub-frame). The backoff interval of the gth random-access slot starts from the time $(t_g+T_{RAR}+W_{RAR}+1)$ and ends at time $(t_g+T_{RAR}+W_{RAR}+W_{BO})$. The MSs transmit their preambles at the ith random-access slot if their backoff counters reach zero within the transmission interval of the ith random-access slot. In the other words, the backoff interval of the gth random-access slot is overlapped with the transmission interval of the ith random-access slot. The transmission interval of the ith random-access slot is the time interval between the $(i-1)$th random-access slot and the ith random-access slot (i.e., $[t_{i-1}+1, t_i]$). $G_{\min}$ is obtained when the right-hand side boundary of the backoff interval of the gth random-access slot reaches the left-hand side boundary of the transmission interval of the ith random-access slot and is given by

$$G_{\min} = (i-1) - \frac{T_{RAR}+W_{RAR}+W_{BO}-1}{T_{RA\_REP}}. \tag{36}$$

$G_{\max}$ is obtained when the left-hand side boundary of the backoff interval of the gth random-access slot exceeds the right-hand side boundary of the transmission interval of the ith random-access slot and is given by

$$G_{\max} = i - \frac{T_{RAR}+W_{RAR}+1}{T_{RA\_REP}}. \tag{37}$$

$\alpha_{g,i}$ is the percentage of the backoff interval of the gth random-access slot that overlaps with the transmission interval of the ith random-access slot. $\alpha_{g,i}$ is given by

$$\alpha_{g,i} = \begin{cases} \frac{t_g+T_{RAR}+W_{RAR}+W_{BO}-t_{i-1}}{W_{BO}}, \\ \text{if } (i-1)-\frac{T_{RAR}+W_{RAR}+W_{BO}-1}{T_{RA\_REP}} \leq g \leq i - \frac{T_{RAR}+W_{RAR}+W_{BO}}{T_{RA\_REP}}, \\ \frac{T_{RA\_REP}}{W_{BO}}, \\ \text{if } i - \frac{T_{RAR}+W_{RAR}+W_{BO}}{T_{RA\_REP}} < g < (i-1) - \frac{T_{RAR}+W_{RAR}}{T_{RA\_REP}}, \\ \frac{t_i-(t_g+T_{RAR}+W_{RAR})}{W_{BO}}, \\ if\ (i-1) - \frac{T_{RAR}+W_{RAR}}{T_{RA\_REP}} \leq g \leq i - \frac{T_{RAR}+W_{RAR}+1}{T_{RA\_REP}}, \\ 0, \text{ otherwise.} \end{cases} \tag{38}$$

Appendix B
Calculation of $\overline{T_n}$

$\overline{T_n}$ is the average access delay of a successfully accessed MS that transmits exactly n preambles. $\overline{T_n}$ contains the time required by the MS to transmit the first preamble, re-transmit the $(n-1)$ preamble(s), wait for the processing of the BS, and finish the message transmission. That is,

$$\overline{T_n} \cong 1 + (n-1)\overline{T_W} + T_{RAR} + W_{RAR} + \overline{T_{MSG}}, \tag{39}$$

where $\overline{T_W}$ is the average waiting time required by an MS to perform backoff and re-transmit a preamble; and $\overline{T_{MSG}}$ is the average message transmission time. Consider the case that an MS transmits a preamble at the 1st random-access slot but fails. Assume that the MS performs random backoff and retransmits a preamble in the $(1+h)$th random-access slot. Therefore, $\overline{T_W}$ is given by

$$\overline{T_W} \cong \sum_{h=H_{\min}}^{H_{\max}} q_h h T_{RA\_REP}, \tag{40}$$

where $H_{\min}$ and $H_{\max}$ are the minimal and the maximal value of h and q_h is the probability of selecting a value of h. $H_{\min}$ and $H_{\max}$ occur when the backoff counter is zero and W_{BO}, respectively. They are given by

$$H_{\min} = \left\lceil \frac{T_{RAR}+W_{RAR}+1}{T_{RA\_REP}} \right\rceil \tag{41}$$

and

$$H_{\max} = \left\lceil \frac{T_{RAR}+W_{RAR}+W_{BO}}{T_{RA\_REP}} \right\rceil. \tag{42}$$

q_h can be determined from the overlapped region of the backoff interval of the 1st random-access slot and the transmission interval of the $(h+1)$th random-access slot. The backoff interval of the first random-access slot starts from the time $(t_1 + T_{RAR} + W_{RAR} + 1)$ and ends at time $(t_1 + T_{RAR} + W_{RAR} + W_{BO})$. The transmission interval of the $(h+1)$th random-access slot starts at $(t_h + 1)$ and ends at t_{h+1}. For $h = H_{\min}$, the overlapped region starts from the left boundary of the backoff interval of the 1st random-access slot and ends at the right boundary of the transmission interval of the $(1+h)$th random-access slot. That is, the overlapped region is $[t_1 + T_{RAR} + W_{RAR} + 1, t_{h+1}]$. For $h = H_{\max}$, the overlapped region starts from the left boundary of the transmission interval of the $(h+1)$th random-access slot and ends at the right-side boundary of the backoff interval of the 1st random-access slot. That is, the overlapped region is $[t_h + 1, t_1 + T_{RAR} + W_{RAR} + 1]$. Therefore, q_h is given by

$$q_h = \begin{cases} \frac{t_{1+h}-(t_1+T_{RAR}+W_{RAR})}{W_{BO}}, & \text{if } h = H_{\min}, \\ \frac{(t_1+T_{RAR}+W_{RAR}+W_{BO})-t_h}{W_{BO}}, & \text{if } h = H_{\max}, \\ \frac{T_{RA\_REP}}{W_{BO}}, & \text{otherwise.} \end{cases} \tag{43}$$

The average message transmission time $\overline{T_{MSG}}$ can be obtained by considering an MS which completes its message transmission by exactly sending u HARQ transmissions of the *Connection Request* message and receiving v HARQ transmissions of the *Connection Response* message $(u, v \le N_{HARQ})$. The time required to transmit u *Connection Request* message and receive v *Connection Response* messages is $(1 + (u-1)(T_{HARQ} + T_{RQ}) + T_{HARQ})$ and $(T_{A\_RP} + (v-1)(T_{HARQ} + T_{RP}) + T_{HARQ})$, respectively. The probability that the message transmission contains u *Connection Request* message and v *Connection Response* messages is ${p_f}^{u-1}(1-p_f){p_f}^{v-1}(1-p_f)$. Hence, $\overline{T_{MSG}}$ is given by

$$\begin{aligned} \overline{T_{MSG}} = & \sum_{u=1}^{N_{HARQ}} \sum_{v=1}^{N_{HARQ}} p_f^{u+v-2}(1-p_f)^2 \left[(1 + (u-1)\right. \\ & \times\ (T_{RQ}+T_{HARQ})+T_{HARQ})+(T_{A\_RP} + (v-1) \\ & \left.\times\ (T_{RP} + T_{HARQ}) + T_{HARQ})\right] \\ = & \sum_{u=1}^{N_{HARQ}} \sum_{v=1}^{N_{HARQ}} p_f^{u+v-2}(1-p_f)^2 \left[(u-1)T_{RQ}\right. \\ & \left. +\ (v-1)T_{RP} + (u+v)T_{HARQ} + T_{A\_RP} + 1\right]. \end{aligned} \tag{44}$$

References

[1] "Evolved universal terrestrial radio access (E-UTRA) medium access control (MAC) protocol specification," Cedex, France, 3GPP TS 36.321 V9.3.0, Jun. 2010.

[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, 2014.

[3] 3GPP TR 37.868, "RAN improvements for machine-type communications," v. 11.0.0, Oct. 2011.

[4] P. Zhou, H. Hu, H. Wang, and H. H. Chen, "An efficient random access scheme for OFDMA systems with implicit message transmission," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2790–2797, Jul. 2008.

[5] S. Kim *et al.*, "Performance evaluation of random access for M2M communication on IEEE 802.16 network," in *Proc. 14th Int. Conf. Adv. Commun. Technol.*, Feb. 2012, pp. 278–283.

[6] R. C. D. Paiva, R. D. Vieira, and M. Saily, "Random access capacity evaluation with synchronized MTC users over wireless networks," in *Proc. IEEE 73th Veh. Technol. Conf.*, May 2012, pp. 1–5.

[7] 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.

[8] L. Kleinrock and F. Tobagi, "Packet switching in radio channels part I Carrier sense multiple-access modes and their throughput delay characteristics," *IEEE Trans. Commun.*, vol. Com-23, no. 12, pp. 1400–1416, Dec. 1975.

[9] A. B. Behroozi-Toosi and R. R. Rao, "Delay upper bounds for a finite user random-access system with bursty arrivals," *IEEE Trans. Commun.*, vol. 40, no. 3, pp. 591–596, Mar. 1992.

[10] Q. Ren and H. Kobayashi, "Transient analysis of media access protocols by diffusion approximation," in *Proc. Dig. Int. Symp. Inf. Theory*, Whistler, BC, Canada, p. 107, Sep. 1995.

[11] D. Shen and V. O. K. Li, "Performance analysis for a stabilized multichannel slotted ALOHA algorithm," in *Proc. 14th IEEE Pers., Indoor Mobile Radio Commun.*, 2003, pp. 591–596.

[12] Y. J. Choi, S. Park, and S. Bahk, "Multichannel random access in OFDMA wireless network," *IEEE J. Sel. Area Commun.*, vol. 24, no. 3, pp. 603–613, Mar. 2006.

[13] R. G. Cheng, C. H. Wei, and S. L. Tsao, "Iterative contending-user estimation method for OFDMA wireless networks with bursty arrivals," in *Proc. 18th IEEE ISCC*, Split, Croatia, Jun. 2013, pp. 000240–000245.

[14] C. H. Wei, R. G. Cheng, and S. L. 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.

[15] "Evolved universal terrestrial radio access (E-UTRA) physical channels and modulation," Cedex, France, 3GPP TS 36.211 V10.2.0, Jun. 2011.

[16] 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.

[17] R. L. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 3, pp. 323–328, May 1987.

[18] F. M. Buckley and P. K. Pollett, "Limit theorems for discrete-time metapopulation models," *Probability Surveys*, vol. 7, p. 5383, 2010.

[19] R. W. R. Darlings and J. R. Norris, "Differential equation approximations for Markov chains," *Probability Surveys*, vol. 5, p. 3779, 2008.

Chia-Hung Wei received the Ph.D. degree in electronic and computer engineering from National Taiwan University of Science and Technology, Taipei, Taiwan, in 2013. He is currently a Senior Engineer with the Smart Network System Institute, Institute for Information Industry, Taipei. His research interests include machine-type communications, multichannel slotted aloha, and performance analysis of 4G cellular networks.

Giuseppe Bianchi has been a Full Professor of networking at the University of Roma Tor Vergata, Rome, Italy, since January 2007. He is documented in about 200 peer-reviewed international journal and conference papers. He has coordinated several large-scale European research projects in the FP6 and FP7 and in the incoming H2020 programs. His research activities include wireless networks, programmable networking, performance evaluation, privacy and security, and traffic monitoring. He has chaired several international networking conferences and workshops, including IEEE Infocom 2014, ACM SRIF 2013, ACM Wintech 2011, and IEEE WoWMoM 2010. He has served as an Associate Editor of the IEEE/ACM TRANSACTIONS ON NETWORKING, an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and an Area Editor of *Elsevier Computer Communication*.

Ray-Guang Cheng (S'94–M'97–SM'07) received the B.E., M.E., and Ph.D. degrees from National Chiao Tung University, Hsinchu, Taiwan, in 1991, 1993, and 1996, respectively, all in communication engineering. From 1997 to 2000, he was a Researcher and a Project Leader with the Advance Technology Center, Computer and Communication Laboratories, Industrial Technology Research Institute (ITRI), Hsinchu. He led the 3G Protocol project and his team was named Top Research Team of the Year by ITRI in 2000. From 2000 to 2003, he was a Senior Manager of the R&D division at BenQ Mobile System Inc., Hsinchu. He is currently a Professor with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. He has authored or coauthored over 90 international journal and conference papers and over 30 IEEE/3GPP standard contributions. He is the holder of IEEE Wireless Communication Professional certification and 18 U.S. patents. His research interests include multihop wireless networks and machine-to-machine communications.

Dr. Cheng is a Senior Member of the IEEE. He was a recipient of the Best Industrial-based Paper Award from the Ministry of Education in 1998, the Advanced Technologies Award from the Ministry of Economic Affairs in 2000, and the Teaching Award, the Research Award, and the Excellence in Counseling Award from NTUST in 2006, 2009, and 2011, respectively. He is also a member of Phi Tau Phi scholastic honor society.