Impact of Retransmission Limit on Preamble Contention in LTE-Advanced Network

Revak R. Tyagi, *Student Member, IEEE*, Frank Aurzada, Ki-Dong Lee, *Senior Member, IEEE*,
Sang G. Kim, *Member, IEEE*, and Martin Reisslein, *Senior Member, IEEE*

Abstract—LTE-Advanced networks employ random access based on preambles transmitted according to multi-channel slotted Aloha principles. The random access is controlled through a limit W on the number of transmission attempts and a time-out period for uniform backoff after a collision. We model the LTE-Advanced random access system by formulating the equilibrium condition for the ratio of the number of requests successful within the permitted number of transmission attempts to those successful in one attempt. We prove that for $W \leq 8$ there is only one equilibrium operating point and for $W \geq 9$ there are three operating points if the request load ρ is between load boundaries ρ_1 and ρ_2. We analytically identify these load boundaries as well as the corresponding system operating points. We analyze the throughput and delay of successful requests at the operating points and validate the analytical results through simulations.

Index Terms—Equilibrium operating points, random access system, slotted Aloha, throughput-delay analysis.

I. INTRODUCTION

LONG Term Evolution (LTE) Advanced (LTE-Advanced) [1] is a popular Radio Access Network (RAN) protocol standard for 4G cellular networks, which has been chosen by many service providers worldwide [2], [3]. Given the prominence of LTE-Advanced in 4G networks it is important to thoroughly analyze its protocol features. In this paper, we focus on the slotted Aloha based random access procedure, which distributed user equipment (UE) nodes must complete to establish a connection to the central evolved node B (eNB). For applications with frequent small data transmissions, such as periodic monitoring of patient vitals in ubiquitous health care systems [4], the random access procedure must be completed for each data transmission, thus efficient and low-delay completion of the random access is highly important.

In brief, the LTE-Advanced random access protocol consists of an access barring check [5], which may bar (block) a UE from attempting to connect to the eNB for a prescribed time period, followed by preamble contention. The preamble contention follows essentially the principles of a multi-channel slotted Aloha system [6] with a limit on the number of retransmissions. Specifically, in a given time slot, the UEs with connection requests randomly select a preamble from among a set of O orthogonal preambles. If two or more UEs select the same preamble, a collision occurs. A UE with a collided preamble retransmits in a later slot, provided it has not exhausted the W permitted transmission attempts. In this paper, we focus on the preamble contention and leave the access barring for future work.

We model the preamble contention for an infinite UE population generating requests according to a Poisson proccess through a system equilibrium condition. We define a function $h(t)$ to represent the ratio of the expected number of requests successfully completed within the W attempts to the number of attempts completed in one attempt for a given expected number of transmitting UEs t. We analyze the equilibrium condition by examining the intersections of the function $h(t)$ with the line t/ρ whose slope is inversely proportional to the load ρ. From this analysis, we show that for $W \leq 8$ there is only a single equilibrium operating point. For $W \geq 9$ we analytically specify load boundaries ρ_1 and ρ_2 that depend only on W, such that for loads ρ in the range (ρ_1, ρ_2), three equilibrium operating points exist, which we analytically specify. For loads ρ outside the $[\rho_1, \rho_2]$ range, only one equilibrium operating point exists for $W \geq 9$. While slotted Aloha systems have been analyzed extensively in the literature (see Section II), to the best of our knowledge, our study is the first to analytically specify the boundaries of the load range giving rise to multiple equilibria for $W \geq 9$ as well as to analytically specify the throughput and delay at these operating points.

The widely expected increase in the number of nodes combined with new services, such as ubiquitous healthcare applications, machine-type and smartphone communication (MTSC) [7], and other small data applications, will frequently generate small data sets. Low delay is often a key requirement for these frequent small data sets, which result in a high random access load. One possible strategy for ensuring low delays is to limit the traffic load. For instance, for a system with transmission attempt limit $W \leq 8$, which has only a single operating point, our delay analysis can be used to limit the load ρ so as to keep the mean delay below a tolerable delay target. For systems with $W \geq 9$, the load can be limited to below the lower load

boundary ρ_1 to avoid the performance degradations due to multiple operating points.

This article is structured as follows. In Section III, we present our model of the LTE-Advanced random access protocol. In Section IV, we formulate the equilibrium condition for the random access system and analyze the preamble contention. In Section V, we prove our results relating to the equilibrium operating points for different transmission limits W, including the load boundaries ρ_1 and ρ_2. In Section VI, we analyze the throughput and delay at the operating point(s). In Section VII, we summarize our findings.

II. RELATED WORK

The throughput-delay performance of slotted Aloha type random access without a limit on the number of transmission attempts, which corresponds to $W \rightarrow \infty$, has been examined in a number of seminal studies for single-channel systems [8]–[15] and multi-channel systems [16]–[20]. For the infinite node model, several of these seminal studies, e.g., [8], [11]–[14], [18] found that for (per channel) loads $\rho < 1/e$, slotted Aloha has three equilibrium operating points, namely one low-delay high-throughput operating point, an intermediate point corresponding to moderate delay and throughput, and a saturation point corresponding to high delay and low throughput; for loads $\rho > 1/e$ (whereby $1/e$ corresponds to our load boundary ρ_2 for large W), only a single saturation point exists. For a multi-channel slotted Aloha system with fast retry (i.e., immediate retransmission in the next slot) or uniform backoff (i.e., retransmission after a uniformly distributed backoff time), we extend these results as follows: We show that for a finite limit on the number of transmission attempts W, $W \geq 9$, there is a load boundary ρ_1 below which only a single operating point exists; for large W, the ρ_1 asymptotically behaves as $(\log W)/W$.

Slotted Aloha based random access with a transmission attempt limit W has been simulated in [21], while the impact of retransmissions on a general packet (cell) queueing system has been analyzed in [22] and a limit of $W = 3$ has been shown to minimize delays in lightly loaded slotted Aloha in [23]. Kwak *et al.* [24] analyzed the effects of limiting the number of transmission attempts on general exponential backoff. Kim [25] formulated an equilibrium condition and observed from exhaustive numerical exploration the existence of either one or three equilibrium operating points for different limits on the number of transmission attempts. Similarly, Liu [26] formulated an equilibrium condition for slotted Aloha with transmission attempt limit and explored this equilibrium condition numerically. In contrast, we formally analyze our equilibrium condition to identify the multiple operating points as well as the corresponding throughputs and delays.

Sakakibara *et al.* [27], [28] and Noguchi *et al.* [29], [30] modeled slotted Aloha systems with the formalisms of catastrophe theory [31], [32]. Through a cusp theory approximation within the catastrophe theory formalism, Sakakibara *et al.* proved that for $W \leq 8$ transmissions there exists only a single equilibrium operating point, whereas for $W \geq 9$ there are load boundaries within which multiple operating points exist. In contrast, we model the underlying slotted Aloha random access dynamics directly through an elementary equilibrium equation. We not

only prove the existence of single and multiple operating points, but also analytically identify the load boundaries that give rise to multiple operating points for $W \geq 9$ as well as analytically identify all operating points. We also analytically derive the throughput-delay performance corresponding to the operating points.

Sarker *et al.* [33]–[36] investigated the impact of limiting the number W of transmission attempts on the throughput. Their work is complementary to ours in that their main focus is on controlling the number of transmission attempts so as to maximize throughput by operating the system near the classical stability limit of $1/e$ of the channel bandwidth (which approximately corresponds to our upper load boundary ρ_2). In contrast to the work of Sarker *et al.*, we include the delay in our evaluations and identify the impact of the single or multiple operating points on throughput and delay.

Recently, the various aspects of LTE random access have attracted significant research interest. Yilmaz *et al.* [37] identified optimization problems for LTE random access. Vukovic and Filipovich [38] examined the impact of different physical random access configurations, such as possible non-uniform distribution of random access opportunities over the slot in the LTE time structure. Kwan and Leung [39] examined the intercell interference caused by neighboring eNBs. Wei *et al.* [40] examined group paging in an LTE network, where each UE in the group has only one requests. Yun [41] comprehensively described the physical (PHY) layer and medium access control (MAC) layer of 3GPP Universal Terrestrial Radio Access, which is related to LTE-Advanced, with a Markov chain model. Our study is complementary to [41] in that we focus on the medium access control and analyze in detail its implications for the existence of a single or multiple operating points, which are not explicitly considered in [41], as well as the resulting throughput and delay. Seo and Leung [42] studied the uniform backoff in LTE relative to the exponential backoff in IEEE 802.16 WiMAX. In the context of this backoff study, Seo and Leung briefly analyzed the implications of these backoff mechanisms on system operating points in the limit $W \rightarrow \infty$ for saturated (high traffic) conditions. Similarly, in [43], Seo and Leung analyzed the initial random access with infinite retransmission limit for LTE semi-persistent scheduling. Our study focused on the impact of the finite transmission attempt limit W and covers the full range of load conditions.

Slotted Aloha based contention with limited number of trials arises also when a mobile reader scans radio frequency identification (RFID) tags [44]. Multiple equilibrium operating points for the mobile RFID reader have been observed by Alcaraz *et al.* [45] and considered in the setting of RFID system parameters. Our analysis complements the Alcaraz *et al.* model in that we analyze the specific underlying conditions that give rise to the multiple operating points and identify these points.

III. MODEL OF LTE-ADVANCED RANDOM ACCESS SYSTEM

A. Random Access Protocol

We consider a single cell in a cellular system, whereby the cell is comprised of a central node, referred to as evolved
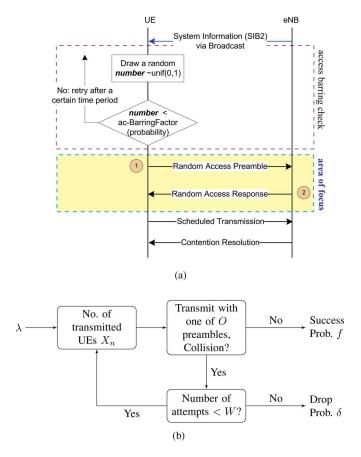
Fig. 1. Illustration of LTE-Advanced Random Access (RA) procedure and the model for the preamble contention: (a) Overall LTE-Advanced random access (RA): The UEs that have passed through the access barring check start transmitting a preamble, and the contention-based transmissions from multiple UEs generate a request load in the random access system. (b) Basic random access model (timeout is considered in Appendix D): Requests for new preamble transmissions arriving at rate λ [requests/slot] and collided transmissions that have not yet exhausted their W permitted attempts contribute to the total number X_n of UEs transmitting a preamble in a slot n. Each transmitting UE randomly selects one of the O preambles.

Node B (eNB), and of multiple User Equipment (UE) nodes. Considering a single cell is not a limitation of our model since the RA procedure in LTE is an interaction between a prescribed UE and its most relevant cell, which is chosen while staying in RRC_IDLE mode or while staying in RRC_CONNECTED mode with time synchronization lost (e.g., when the timing alignment timer expires).

The user equipment (UE) nodes try to establish a communication flow with the eNB, which can be thought of as a circuit-switched connection for the purposes of the present model. (Generally, LTE operates in the packet-switched mode while providing a circuit-switched mode as a fallback; however these details are not relevant for the present model.)

As illustrated in Fig. 1, the overall Random Access (RA) procedure in LTE-Advanced (as well as in the LTE standard preceding LTE-Advanced) consists of an access barring check followed by random access preamble transmission and response. The access barring check allows a UE with probability (access barring factor) P_b to immediately transmit a preamble; while the UE has to wait with probability $1 - P_b$ for an ac-BarringTime. In the present study we focus on the analysis of the preamble transmission by setting $P_b = 1$ and leave the

incorporation of the access barring check into our model for future work.

UEs use one of the Random Access Channel (RACH) opportunities configured by the physical (PHY) layer. The RACH is a set of logical resources defined in the 3-dimensional domain of time-frequency-preamble, whereby the UE randomly chooses a preamble from among O, $O > 1$, allowable preambles.

The eNodeB receives RA requests from UEs during a time slot of duration T_s. If multiple UEs transmitted their requests using the same preamble in the same slot, then those RA requests are considered to have collided. We note that physical layer considerations, such as different levels of transmission power among the UEs, can influence the success or collision of RA requests. The focus of this present study is on capturing the MAC layer behavior and thus detailed physical layer considerations are beyond the scope of this study. When a collision occurs, contention (for RA) is not considered to be resolved, i.e., contention resolution failed. UEs are able to identify the contention resolution result at the fourth step of the RA of LTE [46]. If contention is resolved, the UE enters the RRC_CONNECTED mode.

When contention is not resolved, the UE may repeat the preamble transmission. Specifically, if the UE has had less than W transmission attempts so far, it re-transmits. On the other hand, if the Wth preamble transmission has failed, then the UE drops the request. Before its re-transmission, the UE waits according to a prescribed backoff interval $T_o^{\max}$ ranging from 0 ms to 960 ms, which is signaled by the eNB. For simplicity, we initially set the backoff interval to $T_o^{\max} = 0$ ms, that is, UEs whose preamble transmissions collide in a given slot may re-transmit in the next slot (non-zero backoff intervals are examined in Appendix D). The setting $T_o^{\max} = 0$ corresponds to fast retry in [47]. Each re-transmitting UE uniformly randomly selects a new preamble from among O preambles, independently of the preceding preamble selection.

B. Performance Metrics

The two key performance metrics directly related to the random access procedure are the mean (steady-state) delay D and the mean (steady-state) throughput T of the random access system in equilibrium. We define the delay D of random access as the time period from the instant a UE generates a preamble to the instant the UE is notified about the accepted connection; whereby only requests that are successful within the W transmission attempts are considered in the delay evaluation. We define the throughput T as the long-run average rate at which connection acceptances are granted.

IV. System Equilibrium Analysis

A. Definition of System Characteristics

We model the initial request generation with a Poisson process with a prescribed rate λ [requests/slot]. This model corresponds to an infinite population of "virtual" UEs in the cell, whereby each UE can request a circuit with the eNB.

TABLE I
SUMMARY OF MAIN MODEL NOTATIONS

Slotted Aloha based preamble contention	
O	Number of preambles (equivalent to number of channels in multi-channel slotted Aloha)
T_s	Slot duration for slotted Aloha contention [in seconds]
$T_o^{\max}$	Maximum backoff time [in slots] of uniform backoff
W	Maximum number of transmission attempts
Request traffic model	
λ	Request generation rate [in requests/slot]
$\rho = \frac{\lambda}{O}$	Request load [in requests/slot] per preamble
Random access system model	
X_n	Total number of UEs transmitting a preamble in slot n
ξ_n	Number of UEs transmitting a preamble for a newly generated request in slot n
Z_n	Total number of unsuccessful UEs (with collided preambles) in slot n
f	Probability of successful preamble transmission (without collision) by an UE in a given slot
$\delta = (1-f)^W$	Probability that a UE request collides in all W attempts
x	Expected number of UEs transmitting a preamble (from both new and previously collided requests) in a given slot in steady state
$t = \frac{x}{O}$	Expected number of transmitting UEs per preamble
$y = e^{-t}$	Substitution to simplify notation in system balance equation
$h(t) = \frac{1-\delta}{f} = \frac{1-(1-y)^W}{y}$	Ratio of probability of success within W attempts to probability of success in one attempt
Load boundaries	
$\rho_1,\ \rho_2$	For $W \geq 9$, there is one operating point for loads ρ outside $[\rho_1,\ \rho_2]$; there are three equilibrium operating points if $\rho_1 < \rho < \rho_2$; $\rho_1,\ \rho_2$ depend only on W as per Eqns. (16) and (17)
Performance metrics	
D	Mean delay from request generation to successful contention completion [in slots]
T	Mean throughput of successful contention completions [in request/slot per preamble]

We define X_n to be a random variable denoting the number of UEs that are sending a preamble in a given slot n. We note that both newly generated requests and the re-transmissions of old requests contribute to X_n as analyzed in detail in Section IV-B.

We let ξ_n be a random variable denoting the number of UEs that transmit a preamble for a newly generated request in slot n. For the considered Poisson request arrival process with rate λ, the number of newly generated requests per slot has expected value $E[\xi_n] = \lambda$.

We define f to denote the (steady-state) probability that a UE successfully sends a preamble, i.e., sends a preamble without collision, in a given slot, i.e., any slot in which the UE participates in preamble contention. Note that we model f to be indifferent to the UE's age in retransmission. The probability f is derived in Section IV-C and simulations verifying the model accuracy are presented in Section VI-B.

We define δ to denote the (steady-state) probability that a UE request is unsuccessful in all its W transmission attempts, and as a result drops its request. A UE's attempt in a given slot is unsuccessful with probability $1 - f$, thus the probability that the UE is unsuccessful in all its W attempts can be modeled as

$$\delta = (1-f)^W. \tag{1}$$

Note that $1 - \delta$ is the probability that the UE is successful in one of its (at most W) transmission attempts. The model notation is summarized in Table I.

B. System Balance (Equilibrium) Formulation

We develop a recursion for X_n by noting that the UEs sending a preamble in slot n are either (A) UEs that have generated a new request during the preceding slot and are now

sending their preamble for the first time in slot n, or (B) UEs that experienced a preamble collision in one (or several) preceding slot(s) and have not yet exhausted the maximum number of preamble transmissions W.

Note that X_{n-1} UEs sent a preamble in slot $n-1$; whereby, these UEs either had generated a new request (during slot $n-2$) for preamble transmission and this new request is transmitted for the first time in slot $n-1$, or had a preamble collision in one (or several) preceding slot(s). In steady state, an expected number of $f E[X_{n-1}]$ UEs successfully transmitted a preamble in slot $n-1$. The remaining $(1-f)E[X_{n-1}]$ UEs had a preamble collision and will re-try in slot n, provided they have not exhausted the maximum number of preamble transmissions W. In particular, those UEs that transmitted a preamble for the first time in slot $n-W$ and experienced collisions in all slots $n-W, n-W+1, \ldots, n-2$, and $n-1$ have exhausted their maximum number of preamble transmissions and drop out. Noting that an expected number of $E[\xi_{n-W}]$ UEs had transmitted a preamble for a first time in slot $n-W$, $\delta E[\xi_{n-W}]$ UEs drop out after the preamble contention in slot $n-1$. Thus, there are an expected number of

$$E[X_n] = E[\xi_n] + (1-f)E[X_{n-1}] - \delta E[\xi_{n-W}] \tag{2}$$

UEs transmitting a preamble in slot n. In the illustration in Fig. 2, $E[X_n]$ corresponds to the sum of the left (solid line) portions of $E[\xi_{n-1}], \ldots, E[\xi_{n-W+1}]$, plus all of $E[\xi_n]$. Note that these left portions correspond to $(1-f)E[\xi_{n-1}], \ldots, (1-f)^{W-1}E[\xi_{n-W+1}]$ UEs. Thus, alternatively, we obtain the expected number of UEs transmitting a preamble in slot n as

$$E[X_n] = \sum_{t=0}^{W-1} (1-f)^t E[\xi_{n-t}]. \tag{3}$$

$$\text{\# of UEs w/o success} = \delta E[\xi_{n-W}]$$
$$= (1-f)^W E[\xi_{n-W}] \qquad \text{\# of UEs successful at one of } W \text{ attempts} = (1-\delta)E[\xi_{n-W}]$$

$E[\xi_{n-W}]$

$$(1-f)^{W-1} E[\xi_{n-W}]$$

$E[\xi_{n-W+1}]$

$$\vdots$$

$$\text{\# of UEs collided in slot } n-1 = (1-f)E[\xi_{n-1}] \qquad \text{\# of successful UEs in cycle } n-1 = fE[\xi_{n-1}]$$

$E[\xi_{n-1}]$

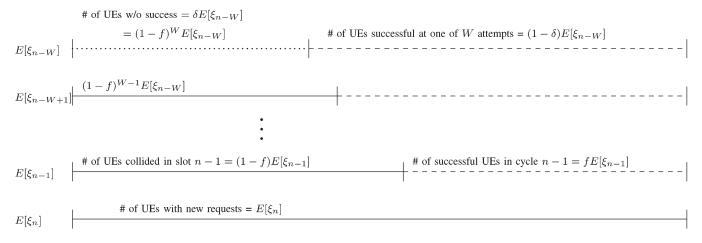$$\text{\# of UEs with new requests} = E[\xi_n]$$

$E[\xi_n]$

Fig. 2. Illustration of dynamics leading to recursion (3) for the expected number of UEs $E[X_n]$ transmitting a preamble in slot n. UEs that have been unsuccessful in slots $n-1$ through $n-W+1$ (represented by the left solid-line portions) retransmit a preamble in slot n. Additionally considering that $E[\xi_n]$ UEs with newly generated requests transmit a preamble in slot n and that $\delta E[\xi_n - W]$ UEs drop after having had no success in W attempts leads to the recursion (2).

Proceeding from (2), we define x to denote the long-run (steady-state) expected value of X_n, noting that in steady state $x = E[X_n] = E[X_{n-1}]$. Thus

$$x = \lambda + (1-f)x - \delta\lambda. \tag{4}$$

Recalling from (1) that $\delta = (1-f)^W$ and rearranging terms gives the steady-state system balance equation

$$\frac{x}{\lambda} = \frac{1-\delta}{f} \tag{5}$$

$$= \frac{1-(1-f)^W}{f}. \tag{6}$$

Intuitively, (5) expresses that the ratio of the expected total number x of transmitting UEs to the expected number λ of UEs transmitting a newly generated request equals the ratio of probability $1-\delta$ of eventual success after at most W attempts to the probability f of success in one attempt. For very low loads, this ratio is one since transmissions are successful in the first attempt $(f \to 1)$ and thus all transmissions are new requests $(x \to \lambda)$. As the load increases, some transmissions fail on the first attempt and lead to an increase in the proportion of retransmissions relative to new transmissions and a commensurate increase in the probability of success after W attempts relative to the success probability in the first attempt. For very high loads, the success probability in a given slot becomes small $(f \to 0)$ and the probability of success after W attempts approaches fW [as $(1-f)^W \approx 1 - fW$ in (6) for small f]. Correspondingly, the expected number of transmitting UEs x approaches the expected number of requests generated in W slots, i.e., λW. Thus, both sides of (5) approach the number of transmission attempts W.

In the following section, we evaluate the probability f of a successful transmission in a slot for the specific preamble transmission procedure in LTE-Advanced. Then, we examine the resulting system balance equation and its implications for system stability.

C. Probability of Successful Preamble Transmission f

Let Z_n be a random variable denoting the total number of unsuccessful UEs in slot n. Note that

$$E[Z_n] = (1-f)x. \tag{7}$$

We denote $\alpha_i, i = 1, \ldots, X_n$, for the preamble randomly selected by UE i. Note that the preambles α_i are independent random variables that are uniformly distributed in $\{1, 2, \ldots, O\}$. A collision occurs if two distinct UEs i and j, $j \neq i$, select the same preamble, i.e.,

$$Z_n = \sum_{i=1}^{X_n} 1_{\{\exists j \in \{1, \ldots, X_n\}, j \neq i: \; \alpha_i = \alpha_j\}}. \tag{8}$$

We evaluate the expectation of the number of unsuccessful UEs Z_n in Appendix A as

$$E[Z_n] \approx x[1 - \mathrm{e}^{-x/O}]. \tag{9}$$

Thus, from (7) and (9)

$$f = \mathrm{e}^{-x/O}. \tag{10}$$

D. Summary

We proceed by inserting (10) in (5). For improved readability we define the preamble load (request generation rate normalized by number of preambles) $\rho := \lambda/O, \rho \geq 0$, and the normalized expected number of UEs transmitting in a slot as $t := x/O, t \geq 0$. The resulting form of the balance equation is

$$\frac{t}{\rho} = \frac{1 - (1 - \mathrm{e}^{-t})^W}{\mathrm{e}^{-t}}. \tag{11}$$

While this nonlinear equation has no closed-form analytical solutions, it can be solved with standard numerical methods. We show in Section V that depending on the values of ρ and W, (11) has one, two, or three solutions for t. From the numerically obtained solutions for t, we obtain the expected numbers of UEs transmitting in a slot as $x = tO$ and the probabilities of
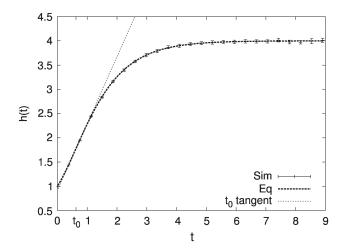
Fig. 3. Comparison of function $h(t)$ from (12) denoted by `Eq` with corresponding simulations (`Sim`) for $W = 4$. The figure also illustrates the tangent to function $h(t)$ at inflexion point t_0 for $W = 4$. Generally, for $W \leq 8$, this tangent crosses the y-axis above the origin and the line t/ρ has a single intersection with $h(t)$ for any load ρ. Thus, only a single equilibrium operating point exists.
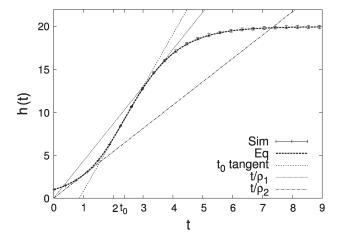

Fig. 4. Comparison of function $h(t)$ from (12) (`Eq`) with simulations (`Sim`) for $W = 20$. The figure also illustrates the tangent to function $h(t)$ at inflexion point t_0 for $W = 20$. Generally, for $W \geq 9$, this tangent crosses the y-axis below the origin and the line t/ρ has a single intersection with $h(t)$ for load $\rho < \rho_1$, three intersections for $\rho_1 < \rho < \rho_2$, and one intersection for $\rho > \rho_2$. Thus, a single, or up to three equilibrium operating points exist and are specified by the intersection(s) of t/ρ with $h(t)$.

successful UE transmission through (10). In order to facilitate the analysis of the balance equation (11), we define for its right-hand side

$$h(t) := g(\mathrm{e}^{-t}) := \frac{1 - (1 - \mathrm{e}^{-t})^W}{\mathrm{e}^{-t}}. \qquad (12)$$

E. Numerical Results

In Figs. 3 and 4 we compare the ratio of the probability of contention success after at most W attempts to the probability of success in one attempt as given by the function $h(t)$ in (12), denoted by `Eq`, with simulations, denoted by `Sim`. (The lines related to t_0, t/ρ_1, and t/ρ_2 in Figs. 3 and 4 are examined in Section V and should be ignored for now.) The simulation model was implemented using OMNeT++ [48] libraries in C++. Statistics collection and execution management was done

using Akaroa2 [49]. In these simulations, we held the number of transmitting UEs t at a prescribed value and observed the mean and 90% confidence interval of the ratio h. We observe from Figs. 3 and 4 that the analytical model for the ratio h given by (12) closely matches the simulation results and thus accurately models the preamble contention.

V. ANALYSIS OF EQUILIBRIUM OPERATING POINTS

A. Preliminaries

The left-hand side of (11) is a line through the origin with slope $1/\rho$. Intersections of the line t/ρ and the function $h(t)$ defined in (12) specify the operating points (equilibrium points) of the system where the balance equation (11) is satisfied.

As shown in Appendix B, $h(t)$ is a strictly increasing function starting at $h(0) = 1$ and ending at $h(\infty) = W$. Furthermore, $h(t)$ has one inflexion point at t_0, whereby the function has one convex piece in the domain $[0, t_0]$ and one concave piece in the domain $[t_0, \infty)$.

B. Single Equilibrium Point for $W \leq 8$

In Appendix B, we show that $h(t)$ has precisely one convex piece (on $[0, t_0]$) and one concave piece (on $[t_0, \infty)$), which implies that the intersection of $h(t)$ and a linear function ($t \mapsto t/\rho$) can have at most three solutions. On the other hand, since $h(0) = 1$ and $h(\infty) = W$, there must be at least one solution.

We now examine the tangent of $h(t)$ at the inflexion point t_0. Note that (11) has three solutions for some ρ if and only if this tangent crosses the y-axis below zero, see Figs. 3 and 4, that is, if and only if

$$h(t_0) - t_0 h'(t_0) < 0. \qquad (13)$$

This equation can readily be checked numerically for any value of W following the equations in Appendix B: calculate the unique solution $z_0 \in (0, 1)$ of $p(z) = 0$ in (40) and then t_0 via (41) and check condition (13). It turns out that condition (13) is violated for all $W \leq 8$ and satisfied for all $W \geq 9$. Thus, for $W \leq 8$ transmission attempts, the balance equation (11) has a single unique solution, i.e., the system has a single equilibrium operating point.

C. Multiple Equilibrium Points for $W \geq 9$

As noted in Section V-B, for all $W \geq 9$, (13) is satisfied, i.e., the tangent on $h(t)$ at the inflexion point t_0 crosses the y-axis below zero, as illustrated in Fig. 4 where the t_0 tangent crosses the x-axis near $t = 1$. Thus, by the piecewise convex and concave property of $h(t)$ shown in Appendix B, there are two tangents on $h(t)$ crossing the origin, illustrated by t/ρ_1 and t/ρ_2 in Fig. 4. These two tangents satisfy

$$h'(t) = \frac{1}{\rho} \quad \text{and} \quad h(t) = \frac{1}{\rho}t \qquad (14)$$

for *some* ρ and t. Substituting $y = \mathrm{e}^{-t}$, these two equations become

$$g'(y)(-y) = \frac{1}{\rho} = \frac{g(y)}{t}, \qquad (15)$$

which is

$$t\left[1 - (1-y)^W - yW(1-y)^{W-1}\right] = 1 - (1-y)^W. \quad (16)$$

Solving (16) for t gives the solutions t_1, t_2, which are those t values where the tangents touch the function $h(t)$. The corresponding slopes $\rho_i, i = 1, 2$, of the tangents are obtained from (14) as $\rho_i = t_i/h(t_i)$, i.e.,

$$\rho_i = \frac{t_i e^{-t_i}}{1 - (1 - e^{-t_i})^W}. \quad (17)$$

Note that ρ_1 and ρ_2 specify the boundaries of the domain of loads ρ where multiple equilibrium operating points exist. In summary, through the analysis in Section V-B and this section, based on the properties of the function $h(t)$ shown in Appendix B, we have proven the following theorem.

Theorem 1: For $W \geq 9$ transmission attempts, there are load boundaries ρ_1, ρ_2, $0 < \rho_1 < \rho_2 < \infty$, that only depend on W according to (16) and (17) such that

1) for $\rho < \rho_1$ the random access system has a single unique equilibrium point;
2) for $\rho = \rho_1$ the random access system has exactly two equilibrium points;
3) for $\rho_1 < \rho < \rho_2$ the random access system has exactly three equilibrium points;
4) for $\rho = \rho_2$ the random access system has exactly two equilibrium points;
5) for $\rho > \rho_2$ the random access system has a single unique equilibrium point.

The one, two, or three equilibrium operating points for a prescribed load ρ are given by the solutions for t of the balance equation (11). If $\rho < \rho_1$ or $\rho > \rho_2$, then, by Theorem 1, the balance equation (11) gives one solution for t; whereas for other ρ values, Theorem 1 states that there are two or three solutions for t. For a given solution t of the balance equation, the corresponding equilibrium operating point in terms of the total expected number x of UEs transmitting a preamble in a slot is given as $x = tO$.

1) Asymptotics for large number of transmission attempts W: We proceed to examine the asymptotics for the load boundaries ρ_1 and ρ_2 as the transmission attempt limit W becomes large. For large W, one solution of (16) is $t_2 \sim 1$, giving

$$\rho_2 \sim e^{-1}, \quad (18)$$

which corresponds to the case $Wy \to \infty$. For the case $Wy \to 0$, we show in Appendix C that

$$\rho_1 \sim \frac{\log \frac{W-1}{2e^{-1}} + \log\log \frac{W-1}{2e^{-1}} - 1}{W} \quad (19)$$

$$\sim \frac{\log W}{W}. \quad (20)$$

2) Numerical Results: In Fig. 5, we plot the load boundaries ρ_1 and ρ_2 as a function of the transmission attempt limit W. We observe that the exact upper boundary ρ_2 from (16) and (17) closely approaches the asymptotic boundary $\rho_2 \sim 1/e$ from (18) even for relatively small W values; for $W \geq 15$, the exact ρ_2 essentially coincides with the asymptotic boundary
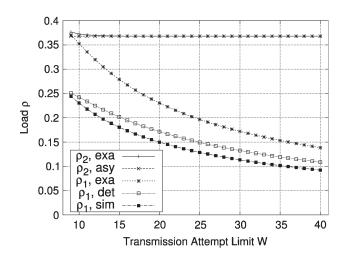

Fig. 5. Lower boundary ρ_1 and upper boundary ρ_2 of load range with multiple equilibrium operating points as a function of number W of permitted transmission attempts. Exact results are obtained with (16) and (17), while detailed (`det`) and simplified (`sim`) asymptotics for ρ_1 are from (19) and (20), respectively, and the asymptotic for ρ_2 is from (18).

$1/e$. On the other hand, the exact lower load boundary ρ_1 from (16) and (17) approaches the asymptotics given by (19) and (20) relatively slowly, with the detailed asymptotic (19) giving a somewhat better approximation for moderately large W values than the simplified asymptotic (20). Thus, for LTE system evaluations, the upper load boundary ρ_2 can be readily approximated by the asymptotic $1/e$ for moderately large W. For the lower boundary ρ_1, the exact analytical characterization through (16) and (17) should be used since the asymptotics overestimate the load range with multiple equilibrium points, especially for small or moderate W values.

We also observe from Fig. 5 that for $W \geq 9$, the width $\rho_2 - \rho_1$ of the load range with multiple equilibrium operating points widens considerably with increasing W, e.g., from $\rho_2 - \rho_1 = 0.1$ for $W = 15$ to 0.2 for $W = 30$. For $W = 200$, the maximum transmission attempt limit in LTE-Advanced, which is not included in Fig. 5 to allow a detailed view of the small W values, ρ_1 drops to 0.0376. That is, multiple operating points exist for loads between 0.0376 and $1/e$ for $W = 200$.

VI. THROUGHPUT-DELAY ANALYSIS

In this section we examine the throughput and delay metrics defined in Section III-B. New requests are generated by the UEs with rate λ [requests/slot], which normalized by the number of preambles O is $\rho = \lambda/O$, and a given UE is successful within the permitted W transmission attempts with probability $1 - \delta$. Thus, the mean throughput of successful requests [requests/slot per preamble] is

$$T = \rho(1 - \delta). \quad (21)$$

A. Delay Analysis

Recall that a preamble transmission is successful with probability f. Thus, the probability of exactly $n, n = 0, 1, \ldots, W - 1$, collisions before a success, can be modeled as $(1-f)^n f$.
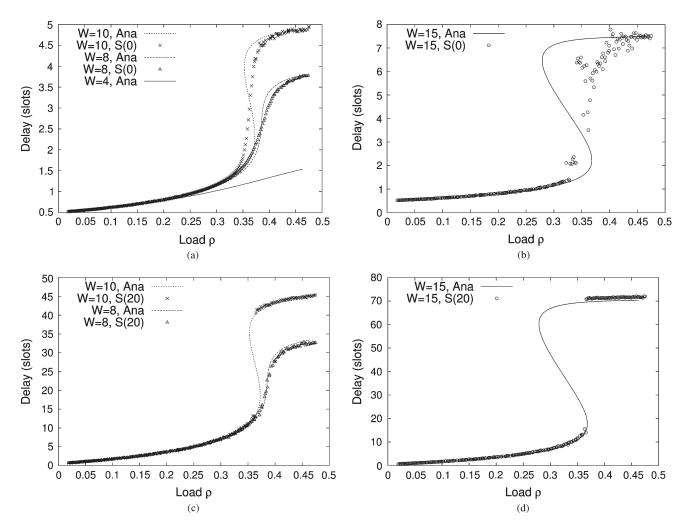
Fig. 6. Mean delay D [in slots] of successful requests, as a function of load ρ [in new requests per slot per preamble] for a range W of transmission attempts. The load boundaries for multiple operating points are $\rho_1 = 0.353$, $\rho_2 = 0.373$ for $W = 10$; while for $W = 15$ they are $\rho_1 = 0.279$, $\rho_2 = 0.368$. For $W \geq 9$ for $\rho_1 < \rho < \rho_2$, we plot the delay values corresponding to the three equilibrium points. $S(T_o^{\max})$ denotes simulation results for a given $T_o^{\max}$ value (a) $W = 4, 8, 10$, $T_o^{\max} = 0$ (b) $W = 15$, $T_o^{\max} = 0$ (c) $W = 4, 8, 10$, $T_o^{\max} = 20$ (d) $W = 15$, $T_o^{\max} = 20$.

Hence, the probability of a UE to experience n collisions, given that it sends (i.e., experiences any number k, $k = 0, 1 \ldots, W - 1$ collisions) is

$$\frac{(1 - f)^n f}{\sum_{k=0}^{W-1} (1 - f)^k f}. \tag{22}$$

Each collision increases the delay by one slot of duration T_s. Thus, the expected delay due to collisions is

$$D_c = T_s \cdot \sum_{n=0}^{W-1} n \cdot \frac{(1 - f)^n f}{\sum_{k=0}^{W-1} (1 - f)^k f}. \tag{23}$$

We model the delay from the instant of request generation to the next time slot boundary (backward recurrence time) [50] with the additive constant $T_s/2$. We further employ the summation identity for $0 < y < 1$

$$\sum_{k=0}^{W-1} k \cdot y^k = y \cdot \frac{1 + (W - 1)y^W - W y^{W-1}}{(1 - y)^2}. \tag{24}$$

Hence, for the preamble transmission success probability f obtained through (11) and (10)

$$D = T_s \left(\frac{1}{f} - 1 \right) \frac{1 + (W - 1)(1 - f)^W - W(1 - f)^{W-1}}{1 - (1 - f)^W}$$

$$+ \frac{T_s}{2}. \tag{25}$$

With uniform backoff with $T_o^{\max}$, as outlined in Appendix D, the average delay caused by a collision increases from T_s to $(1 + (T_o^{\max}/2))T_s$, i.e., T_s has to be replaced by $(1 + (T_o^{\max}/2))T_s$ in the first summand of (25).

B. Evaluation Results

In Figs. 6 and 7, we plot the mean delay [in slots] experienced by a successful request and the throughput $T = \rho(1 - \delta)$ of successful requests per slot per preamble as a function of the load ρ. For a relatively small transmission attempt limit $W = 4$ without backoff, i.e., $T_o^{\max} = 0$, we observe that the successful requests experience a low delay of less than $D = 1.17$ slots at load $\rho = 0.35$, see Fig. 6(a), while Fig. 7(a) indicates that the
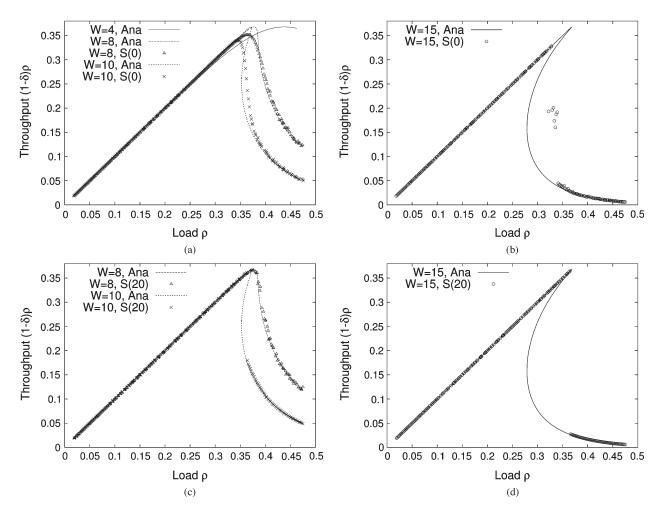
Fig. 7. Throughput $T = \rho(1 - \delta)$ [in successful requests per slot per preamble] as a function of load ρ [in new requests per slot per preamble] for a range W of transmission attempts. The load boundaries for multiple operating points for $W = 10$ are $\rho_1 = 0.353$, $\rho_2 = 0.373$, while the boundaries for $W = 15$ are $\rho_1 = 0.279$, $\rho_2 = 0.368$. Simulation results for $T_o^{\max} = 0$ are denoted by S(0), while S(20) denotes simulation results for $T_o^{\max} = 20$ slots (a) $W = 4, 8, 10$, $T_o^{\max} = 0$ (b) $W = 15$, $T_o^{\max} = 0$ (c) $W = 4, 8, 10$, $T_o^{\max} = 20$ (d) $W = 15$, $T_o^{\max} = 20$.

throughput is $T = 0.334$. For $W = 8$, the highest transmission attempt limit that guarantees a single operating point for all loads ρ, without backoff, the delays are moderately higher with $D = 1.49$ slots while the throughput is very slightly higher $T = 0.348$ at load $\rho = 0.35$ compared to $W = 4$. At this moderately high load level, having the unsuccessful transmissions attempt more re-transmissions is beneficial in that it slightly increases the throughput, while only moderately increasing the delay.

As the load increases beyond 0.37, we observe from Figs. 6 and 7 that the contention with $W = 8$ leads to rapidly increasing delays while the throughput drops sharply. In contrast, for $W = 4$, the system degrades more gracefully, with the mean delay starting to level out around $D = 1.8$ slots for load levels of $\rho = 0.6$ (i.e., outside the plotted range) while the throughput drops to $T = 0.1$ for a load of $\rho = 0.925$. For high loads, the success probability f becomes small($f \to 0$), and the mean delay due to collisions D_c given by (23) correspondingly approaches $\lim_{f \to 0} D_c = (W - 1)T_s/2$; adding the mean waiting time $T_s/2$ from request generation to next slot boundary gives a maximum expected delay of $WT_s/2$. Intuitively, for small success probability f, almost all requests undergo W transmission attempts (and nearly all requests drop after their

Wth attempt). The few requests that are successful, experience their success after a number of attempts that is approximately uniformly distributed over $[1, W]$ (corresponding to $[0, W - 1]$ experienced collisions).

The sharp throughput drop at very high loads, e.g., load $\rho = 0.45$, for $W = 8$ compared to $W = 4$ observed in Fig. 7(a) is mainly due to the exacerbation of the overload conditions by the higher number of transmission attempts with $W = 8$. Specifically, for load $\rho = 0.45$, we found from the numerical evaluation of the analysis in Section IV that for $W = 4$, there are on average $x = 57.5$ preambles transmitted per slot with a steady-state success probability $f = 0.345$ resulting in a throughput of $T = 0.367$. In contrast, with $W = 8$ transmission attempts, there are $x = 165.4$ transmitted preambles with success probability $f = 0.047$ and throughput $T = 0.143$. At this particular high load level, the doubled number of transmission attempts with $W = 8$ roughly triples the number of contending transmissions in each slot, which reduces the success probability to roughly one seventh of the success probability for $W = 4$. The higher number of contending requests with $W = 8$ somewhat compensates for this dramatically lower success probability, but the throughput with $W = 8$ is still less than half compared to $W = 4$.

TABLE II
COMPARISON OF FINITE UE SIMULATION MODEL (N UEs, IDLE UE GENERATES NEW REQUEST WITH PROBABILITY p IN A SLOT) WITH SIMULATION AND ANALYSIS INFINITE UE MODEL ($N \to \infty$, POISSON REQUEST GENERATION RATE $\lambda = Np$ REQ./SLOT) FOR $W = 4$

N	$\rho = 0.30$		$\rho = 0.35$		$\rho = 0.40$		$\rho = 0.45$	
	D	T	D	T	D	T	D	T
100 (sim)	0.96	0.30	1.05	0.34	1.15	0.38	1.22	0.43
1000 (sim)	1.02	0.29	1.15	0.33	1.30	0.36	1.44	0.38
10000 (sim)	1.02	0.29	1.17	0.33	1.33	0.36	1.47	0.37
∞ (sim)	1.03	0.29	1.16	0.33	1.33	0.36	1.48	0.36
∞ (ana)	1.02	0.29	1.17	0.33	1.33	0.36	1.50	0.37

For $T_o^{\max} = 20$ slots for $W = 4$ and 8, we observe a close to ten-fold increase of the mean delays in Fig. 6(c) and (d) compared to Fig. 6(a) and (b), while the throughput remains essentially unchanged in Fig. 7. The $T_o^{\max} = 20$ slot timeout increases the delay introduced by a collision from one slot to on average $(1 + T_o^{\max}/2) = 11$ slots [see discussion immediately following (25)]. On the other hand, as outlined in Appendix D, a static timeout $T_o^{\max}$ does not affect the steady-state drop probability δ and thus preserves the steady-state throughput.

In the simulations for this section, the Poisson generation rate λ of new requests was incremented with a step size of 0.1, corresponding to a step size for $\rho = \lambda/O$ of 0.00185 for $O = 54$ preambles, which is a typical operational O value for LTE-Advanced networks. Each point represents the mean of a simulation run long enough such that the 90% confidence intervals of both performance metrics are less than 10% of their corresponding sample means. We observe from Figs. 6 and 7 that the analytical model closely approximates the simulation results for $W = 8$. The simulation results for $W = 4$, which match very closely to the plotted analysis results, were omitted to avoid clutter.

In Table II, we compare the Poisson request generation model with rate λ, which represents an infinite UE population, with a corresponding simulation model for a finite number of N UEs, whereby each of the N UEs generates a new request only when it is idle with probability p in a slot. We observe that compared to the Poisson model, the finite UE model gives smaller delays D and higher throughputs T as the number N of UEs decreases and the load ρ increases. For smaller N and higher ρ relatively more of the UEs are backlogged with a collided request that is being retransmitted, thus reducing the effective request generation rate. The Poisson model represents a worst-case request generation model in that the generation rate of new requests stays constant, irrespective of the number of backlogged requests.

Turning to the results for $W = 10$ and 15, we observe that the analytical model and simulations closely match for loads outside the (ρ_1, ρ_2) range. For loads inside the ρ_1 to ρ_2 range, a plotted simulation point for a given load gives the mean of the respective performance metric (delay or throughput) experienced in a very long simulation run. That is, the simulation point reflects the multiple operating points and the delays and throughputs experienced at these operating points weighted by the sojourn times at these operating points.

We turn to the effect of backoff for $W = 15$. We observe from the simulation results that the uniform backoff with $T_o^{\max} = 20$ slots helps to achieve essentially zero drop probability and

consequently throughput equal to the traffic load for loads up to approximately 0.361 in Fig. 7(d), whereas without backoff ($T_o^{\max} = 0$), drop probabilities of close to zero occur only for loads up to around 0.320 in Fig. 7(b). (In Fig. 6(b) and (d), these load values correspond to the loads where the delays "jump up" from the respective lower segments of the S-shaped delay curves.) The backoff uniformly redistributes the collided UEs from a given slot that have not exhausted their W attempts over the following $T_o^{\max} + 1$ slots. This redistribution effectively "smoothes" the number of UEs rejoining the contention and lowers the probability of the system entering the operating points with higher drop probabilities and delays. Note that this smoothing effect comes at the expense of greatly increased mean delay. For instance, for a load of $\rho = 0.31$, the mean delay is 7.77 slots with $T_o^{\max} = 20$ slots compared to a mean delay of 1.29 slots with $T_o^{\max} = 0$. We note that the results in Fig. 7 indicating relatively small benefits of uniform backoff for contention with a typical number of $O = 54$ preambles are complementary to the results displayed in [42, Fig. 2], which considers the special case of $O = 1$ preamble. The results in [42, Fig. 2] indicate significant throughput increases albeit at the expense of substantially increased delay due to uniform backoff. With uniform backoff, the collided UEs from a given slot are effectively randomly redistributed to the $O \cdot (T_o^{\max} + 1)$ preambles occurring over the next $T_o^{\max} + 1$ slots. Thus, for a very small number O of preambles, the uniform backoff can help in reducing future collisions, thus increasing throughput. For the typical, moderately large numbers on the order of tens of preambles, the effect of backoff diminishes, as observed in Fig. 6.

Turning to the comparison of the performance for $W = 10$ and 15 with the $W = 4$ and 8 values without multiple operating points, we observe from Fig. 6(a) that the mean delay for $W = 10$ at its $\rho_1 = 0.353$ load is approximately 1.88 slots compared to 1.61 slots for $W = 8$ at the 0.353 load. The mean delays for $W = 15$ and $W = 8$ for $T_o^{\max} = 0$ at $\rho_1 = 0.279$ (for $W = 15$) are essentially the same 1.03 slots. Notice also that for $W = 10$ and 15, the throughput is close to the request arrival rate for loads $\rho < \rho_1$. We furthermore observe that for $W = 10$ and 15 with $T_o^{\max} = 0$, the performance can degrade quite considerably for loads $\rho > \rho_1$, especially toward the middle and upper end of the ρ_1 to ρ_2 range.

VII. CONCLUSION

We have analyzed the impact of the number W of transmission attempts on the throughput and delay of the slotted Aloha based preamble contention in the LTE-Advanced random access system. Our study provides analytical characterizations for the combinations of transmission attempt limit W and request load ρ that results in one, two, or three equilibrium operating points. Specifically, for $W \geq 9$ transmission attempts, which are a necessary condition for multiple operating points, we analyze the load region (ρ_1, ρ_2) that results in three operating points. We analytically characterize the throughputs and delays at these operating points.

The numerical investigations with our analysis results and verifying simulations indicate that for the examined scenario

with $O = 54$ preambles, a small to moderately large transmission attempt limit W around ten without backoff gives good throughput-delay performance. Uniform backoff achieves only relatively small throughput improvements at the expense of substantially increased delays. For reliable low-delay service, a network with $W \geq 9$ should be operated with a load below the boundary ρ_1, which ensures that the network does not experience high-delay operating points. For $W \leq 8$, our delay analysis can be used to identify load limits for low-delay service.

There are many important directions for future research. One example direction is to examine service differentiation [51], [52] whereby different service classes employ different transmission attempt limits W. Another direction is to study the Internetworking of LTE-Advanced networks with local networks, such as body area networks, attached to the UE and backhaul networks, such as Ethernet passive optical networks [53], [54] or metropolitan area optical networks [55]–[57], attached to the eNB.

APPENDIX A
EXPECTED NUMBER OF UNSUCCESSFUL UEs $E[Z_n]$

Proceeding from (8), we evaluate the conditional expectation of the number of unsuccessful UEs Z_n given the number of UEs X_n sending a preamble in slot n, as

$$E[Z_n|X_n] = \sum_{i=1}^{X_n} E[1_{\{\exists j \in \{1,\ldots,X_n\}, j \neq i: \alpha_i = \alpha_j\}}] \quad (26)$$

$$= \sum_{i=1}^{X_n} [1 - P(\forall j \in \{1,\ldots,X_n\}, j \neq i : \alpha_i \neq \alpha_j)] \quad (27)$$

$$= \sum_{i=1}^{X_n} \left[1 - \prod_{j=1, j \neq i}^{X_n} P(\alpha_i \neq \alpha_j)\right] \quad (28)$$

$$= X_n \left(1 - \left(\frac{O-1}{O}\right)^{X_n-1}\right) \quad (29)$$

whereby in the last step we substituted $P(\alpha_i \neq \alpha_j) = (O-1)/O$ as there are $O - 1$ preambles (out of the total of O preambles) that are not equal to a given (fixed) preamble and the UEs select the preambles independently. We note that X_n follows approximately a Poisson distribution (with mean x). To see this, observe from (3) and the illustration in Fig. 2 that X_n is a sum of random fractions of Poisson random variables. From (29), we evaluate $E[Z_n] = E[E[Z_n|X_n]]$ as follows:

$$E[Z_n] = E\left[X_n \left(1 - \left(\frac{O-1}{O}\right)^{X_n-1}\right)\right] \quad (30)$$

$$\approx \sum_{k=1}^{\infty} \frac{x^k}{k!} e^{-x} \cdot k \left(1 - \left(\frac{O-1}{O}\right)^{k-1}\right) \quad (31)$$

$$= x\left[1 - e^{-x} \exp\left(x\frac{O-1}{O}\right) \sum_{k=0}^{\infty} \frac{1}{k!} \left(x\frac{O-1}{O}\right)^k\right.$$

$$\left. \times \exp\left(-x\frac{O-1}{O}\right)\right] \quad (32)$$

$$= x[1 - e^{-x/O}] \quad (33)$$

whereby the summation in (32) is over the probability mass function of a Poisson random variable with mean $x(O-1)/O$, i.e., gives one.

APPENDIX B
PROPERTIES OF $h(t)$

In this appendix, we analyze the right-hand side of the balance equation (11), i.e.,

$$h(t) := g(e^{-t}) := \frac{1 - (1 - e^{-t})^W}{e^{-t}} \quad (34)$$

which represents the ratio of steady-state success probability after W transmission attempts to success probability in the first attempt. We readily verify that $h'(t) = 0$ has no solutions: Indeed, abbreviating $y = e^{-t} \in (0, 1]$ we have

$$h'(t) = g'(e^{-t})(-e^{-t}). \quad (35)$$

Clearly

$$g'(y) = \frac{d}{dy} \frac{1 - (1-y)^W}{y} \quad (36)$$

$$= \frac{W(1-y)^{W-1}y - [1 - (1-y)^W]}{y^2}. \quad (37)$$

The numerator of (37) has no zeros for $y \in (0, 1)$. In fact, this numerator is negative for all $y \in (0, 1)$. Thus h' is positive for all $t > 0$, showing that h is a strictly increasing function starting at $h(0) = 1$ and ending at $h(\infty) = W$.

We next show that $h(t)$ has precisely one convex and one concave piece. Specifically, we show that the equation $h''(t) = 0$ has exactly one solution for $t > 0$, which we will call inflexion point t_0. This shows that h has exactly one convex piece (for arguments in $[0, t_0]$) and one concave piece (for the arguments in $[t_0, \infty)$).

In order to analyze the equation $h''(t) = 0$, note that

$$h'(t) = g'(y)(-y), \quad h''(t) = g''(y)(-y)^2 + g'(y)y. \quad (38)$$

So that $h''(t) = 0$ has a solution for $t > 0$ if and only if

$$g''(y)y + g'(y) = 0 \quad (39)$$

has a solution for $y \in (0, 1)$. After some simplifications, and setting $z := 1 - y$, this is equivalent to

$$0 = p(z) := 1 - W(W-1)z^{W-2} + W(2W-3)z^{W-1} - (W-1)^2 z^W. \quad (40)$$

This is now a polynomial in z of degree W. It can be seen easily that $p(0) = 1$, $p(1) = 0$, $p'(0) = 0$, $p'(1) = 0$, $p'((W-2)/(W-1)) = 0$, and $p''(1) = -W(W-1) < 0$. Thus, $p(z) = 0$ has exactly one solution in $(0, 1)$ (which we denote by z_0) and, going back, so has $h''(t) = 0$, at

$$t_0 := \log \frac{1}{1 - z_0}. \quad (41)$$

APPENDIX C
ASYMPTOTICS OF ρ_1 FOR LARGE W

In this appendix, we derive the asymptotics of the load boundary ρ_1 for large transmission attempt limits W (with $Wy \to 0$) given in (19). We first approximate (17) with the Taylor expansion

$$(1-y)^W = 1 - Wy + \frac{W(W-1)}{2}y^2 + \mathcal{O}\left((Wy)^3\right) \quad (42)$$

to obtain

$$\rho_1 = \frac{t_1 e^{-t_1}}{1 - [1 - We^{-t_1} + \mathcal{O}\left((We^{-t_1})^2\right)]} \quad (43)$$

$$\sim \frac{t_1}{W}. \quad (44)$$

We proceed to express t_1 asymptotically in terms of W, whereby we omit in the following the subscript 1 to avoid clutter. We approximate (16) with the Taylor expansions (42) and

$$Wy(1-y)^{W-1} = Wy\left[1 - (W-1)y\right] + \mathcal{O}\left((Wy)^3\right) \quad (45)$$

to obtain after algebraic simplifications and recalling $y = e^{-t}$ that

$$-(t+1)e^{-(t+1)} = -\frac{2e^{-1}}{W-1}. \quad (46)$$

The solution $t = t(W)$ of (46) is a Product-log function, also referred to as Lambert W function [58], whereby the W in the Lambert W function is not to be confused with our notation W for the transmission attempt limit. For the asymptotic behavior of this function, specifically its branch -1, it can be shown that

$$t(W) = \log\frac{W-1}{2e^{-1}} + \log\log\frac{W-1}{2e^{-1}} - 1 + o(1) \quad (47)$$

where $o(1)$ denotes a term that tends to zero as $W \to \infty$. Inserting (47) in (44) gives (19).

An alternative approach to employing the Lambert W function is to express $t = t(W)$ by defining $\omega = (W-1)/(2e^{-1})$ and

$$t + 1 = \log[\omega s] \quad (48)$$

with s to be determined. Then, inserting (48) in the dominating exponential term $e^{-(t+1)}$, (46) becomes

$$(t+1)\frac{1}{\omega s} = \frac{1}{\omega} \quad (49)$$

i.e., $t + 1 = s$ and thus

$$s = t + 1 = \log[\omega s] \quad (50)$$

$$= \log[\omega] + \log[s] \quad (51)$$

$$= \log[\omega] + \log[t+1] \quad (52)$$

$$= \log\omega + \log\log[\omega s] \quad (53)$$

$$= \log\omega + \log\log\omega + o(1). \quad (54)$$

This leads to

$$t = s - 1 = \log\omega + \log\log\omega - 1 + o(1), \quad (55)$$

which is equivalent to (47).

APPENDIX D
IMPACT OF UNIFORM BACKOFF WITH $T_o^{\max}$

In this section, we outline the impact of uniform backoff with backoff interval (maximum backoff time) $T_o^{\max}$. Suppose that the $T_o^{\max}$ is given as a static value in units of the slot time T_s. Then, a UE that has suffered a preamble collision and has not yet exhausted its W transmission attempts draws a backoff time uniformly randomly from $[0, T_o^{\max}]$. Thus, preamble transmissions that collided in the present slot (and have not exhausted their W attempts) are effectively randomly distributed to rejoin the contention over the next $T_o^{\max} + 1$ slots. Specifically, on average $1/(T_o^{\max} + 1)$ of the collided UEs are distributed to each of the next $T_o^{\max} + 1$ slots. From the analysis leading up to (2), we see that for a given slot $n - 1$ with an expected number of x transmitting UEs, an expected number of $(1 - f)x$ experience a collision, out of which an expected number of $\delta\lambda$ have exhausted their W attempts (after their attempt in slot $n - 1$) and drop out. Thus, there are an expected number of $(1 - f)x - \delta\lambda$ UEs that retransmit in a future slot $n, n + 1, \ldots, n + T_o^{\max}$.

Now consider one of these future slots, say slot n. A fraction $1/(T_o^{\max} + 1)$ of the collided UEs (that have not yet exhausted their W attempts) from each of the preceding $T_o^{\max} + 1$ slots, namely slots $n - 1, n - 2, \ldots, n - (T_o^{\max} + 1)$, rejoins the contention in slot n. Thus, there are

$$(T_o^{\max} + 1)\frac{1}{T_o^{\max} + 1} \cdot [(1-f)x - \delta\lambda] = (1-f)x - \delta\lambda \quad (56)$$

UEs rejoining the contention, leaving the balance equation (2), which governs the equilibrium operating points, unchanged. Consequently, the steady-state drop probability δ in our model is not affected by static $T_o^{\max}$. Modeling adaptive control strategies with dynamic $T_o^{\max}$ is an important direction for future research.

REFERENCES

[1] A. Larmo, M. Lindstrom, M. Meyer, G. Pelletier, J. Torsner, and H. Wiemann, "The LTE link-layer design," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 52–59, Apr. 2009.

[2] *4G Mobile Broadband Evolution: 3GPP Release 10 and Beyond, HSPA +, SAE/LET and LTE-Advanced*, 4G Americas, Bellevue, WA, USA, Jan. 2012.

[3] N. Boudriga, M. Obaidat, and F. Zarai, "Intelligent network functionalities in wireless 4G networks: Integration scheme and simulation analysis," *Comp. Commun.*, vol. 31, no. 16, pp. 3752–3759, Oct. 2008.

[4] K.-D. Lee and A. Vasilakos, "Access stratum resource management for reliable u-healthcare service in LTE networks," *Wireless Netw.*, vol. 17, no. 7, pp. 1667–1678, Oct. 2011.

[5] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 27–32, Jan. 2012.

[6] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis*. Berlin, Germany: Springer-Verlag, 1989.

[7] S. Lien, K. 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.

[8] A. Carleial and M. Hellman, "Bistable behavior of ALOHA-type systems," *IEEE Trans. Commun.*, vol. COM-23, no. 4, pp. 401–410, Apr. 1975.

[9] L. Dai, "Stability and delay analysis of buffered Aloha networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2707–2719, Aug. 2012.

[10] M. Ferguson, "On the control, stability, and waiting time in a slotted ALOHA random-access system," *IEEE Trans. Commun.*, vol. COM-23, no. 11, pp. 1306–1311, Nov. 1975.

[11] Y.-C. Jenq, "On the stability of slotted ALOHA systems," *IEEE Trans. Commun.*, vol. COM-28, no. 11, pp. 1936–1939, Nov. 1980.

[12] S. Kamal and S. Mamoud, "A study of users' buffer variations in random access satellite channels," *IEEE Trans. Commun.*, vol. COM-27, no. 6, pp. 857–868, Jun. 1979.

[13] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Performance evaluation," *IEEE Trans. Commun.*, vol. COM-23, no. 4, pp. 410–423, Apr. 1975.

[14] R. Murali and B. Hughes, "Coding and stability in frequency-hop packet radio networks," *IEEE Trans. Comm.*, vol. 46, no. 2, pp. 191–199, Feb. 1998.

[15] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted ALOHA with multipacket reception," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2636–2656, Jul. 2005.

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

[17] D. Shen and V. Li, "Stabilized multi-channel ALOHA for wireless OFDM networks," in *Proc. IEEE GLOBECOM*, 2002, pp. 701–705.

[18] W. Szpankowski, "Packet switching in multiple radio channels: Analysis and stability of a random access system," *Comput. Netw.*, vol. 7, no. 1, pp. 17–26, Feb. 1983.

[19] W. Yue, "The effect of capture on performance of multichannel slotted ALOHA systems," *IEEE Trans. Commun.*, vol. 39, no. 6, pp. 818–822, Jun. 1991.

[20] Z. Zhang and Y.-J. Liu, "Comments on 'the effect of capture on performance of multichannel slotted ALOHA systems'," *IEEE Trans. Commun.*, vol. 41, no. 10, pp. 1433–1435, Oct. 1993.

[21] C. Lüders and R. Haferbeck, "The performance of the GSM random access procedure," in *Proc. IEEE Veh. Technol. Conf.*, Jun. 1994, vol. 2, pp. 1165–1169.

[22] S. Grishechkin, M. Devetsikiotis, I. Lambadaris, and C. Hobbs, "Multistability in queues with retransmission and its relationship with large deviations in branching processes," *Theory Probab. Appl.*, vol. 47, no. 1, pp. 139–150, 2003.

[23] B. Simon and L. Votta, "The optimal retry distribution for lightly loaded slotted Aloha systems," *IEEE Trans. Commun.*, vol. COM-33, no. 7, pp. 724–725, Jul. 1985.

[24] B.-J. Kwak, N.-O. Song, and L. Miller, "Performance analysis of exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 343–355, Apr. 2005.

[25] S. W. Kim, "Frequency-hopped spread-spectrum random access with retransmission cutoff and code rate adjustment," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 2, pp. 344–349, Feb. 1992.

[26] Y.-S. Liu, "Performance analysis of frequency-hop packet radio networks with generalized retransmission backoff," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 703–711, Oct. 2002.

[27] K. Sakakibara, H. Muta, and Y. Yuba, "The effect of limiting the number of retransmission trials on the stability of slotted ALOHA systems," *IEEE Trans. Veh. Technol.*, vol. 49, no. 4, pp. 1449–1453, Jul. 2000.

[28] K. Sakakibara, T. Seto, D. Yoshimura, and J. Yamakita, "Effect of exponential backoff scheme and retransmission cutoff on the stability of frequency-hopping slotted ALOHA systems," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 714–722, Jul. 2003.

[29] Y. Onozato, J. Liu, S. Shimamoto, and S. Noguchi, "Effect of propogation delays on ALOHA systems," *Comp. Netw.*, vol. 12, no. 5, pp. 329–337, 1986.

[30] X. Wang, J. Kaniyil, Y. Onozato, and S. Noguchi, "Heterogeneous ALOHA networks: A sufficient condition for all equilibrium states to be stable," *Comput. Netw. ISDN Syst.*, vol. 22, no. 3, pp. 213–224, Oct. 1991.

[31] Y. Onozato and S. Noguchi, "A unified analysis of steady state behavior in random access schemes," *Comput. Netw. ISDN Syst.*, vol. 10, no. 2, pp. 111–122, Sep. 1985.

[32] Y. Onozato and S. Noguchi, "On the thrashing cusp in slotted Aloha systems," *IEEE Trans. Commun.*, vol. COM-33, no. 11, pp. 1171–1182, Nov. 1985.

[33] J. H. Sarker and S. J. Halme, "An optimum retransmission cut-off scheme for slotted ALOHA," *Wireless Pers. Commun.*, vol. 13, no. 1/2, pp. 185–202, May 2000.

[34] J. H. Sarker, "Stable and unstable operating regions of slotted ALOHA with number of retransmission attempts and number of power levels," *Proc. Inst. Elect. Eng.—Commun.*, vol. 153, no. 3, pp. 355–364, Jun. 2006.

[35] J. H. Sarker and H. T. Mouftah, "Stability of multiple receiving nodes slotted ALOHA for wireless ad hoc networks," in *Proc. IEEE GLOBECOM*, 2008, pp. 1–5.

[36] J. Sarker and H. Mouftah, "Self-stability of slotted ALOHA by limiting the number of retransmission trials in infrastructure-less wireless networks," *Telecommun. Syst.*, vol. 52, no. 2, pp. 435–444, Feb. 2013.

[37] O. Yilmaz, J. Hamalainen, and S. Hamalainen, "Self-optimization of random access channel in 3rd generation partnership project long term evolution," *Wireless Commun. Mobile Comput.*, vol. 11, no. 12, pp. 1507–1517, Nov. 2011.

[38] I. Vukovic and I. Filipovich, "Throughput analysis of TDD LTE random access channel," in *Proc. Int. Symp. PIMRC*, Sep. 2011, pp. 1652–1656.

[39] R. Kwan and C. Leung, "On collision probabilities in frequency-domain scheduling for LTE cellular networks," *IEEE Commun. Lett.*, vol. 15, no. 9, pp. 965–967, Sep. 2011.

[40] 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.

[41] J. Yun, "Cross-layer analysis of the random access mechanism in universal terrestrial radio access," *Comput. Netw.*, vol. 56, no. 1, pp. 315–328, Jan. 2012.

[42] J. Seo and V. Leung, "Design and analysis of backoff algorithms for random access channel in UMTS-LTE and IEEE 802.16 system," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.

[43] J. Seo and V. Leung, "Performance modeling and stability of semi-persistent scheduling with initial random access in LTE," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4446–4456, Dec. 2012.

[44] V. Sarangan, M. Devarapalli, and S. Radhakrishnan, "A framework for fast RFID tag reading in static and mobile environments," *Comput. Netw.*, vol. 52, no. 5, pp. 1058–1073, Apr. 2008.

[45] J. J. Alcaraz, E. Egea-Lopez, J. Vales-Alonso, and J. Garcia-Haro, "Dynamic system model for optimal configuration of mobile RFID systems," *Comput. Netw.*, vol. 55, no. 1, pp. 74–83, Jan. 2011.

[46] "Evolved universal terrestrial radio access (E-UTRA); Medium access control (MAC) protocol spec.," Sophia-Antipolis, France, 3GPP TS 36.321 v.10.4.0, 2011.

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

[48] A. Varga, "The OMNeT++ discrete event simulation system," in *Proc. ESM*, Jun. 2001, pp. 319–324.

[49] G. Erwing, K. Pawlikowski, and D. McNickle, "Akaroa2: Exploiting network computing by distributing stochastic simulation," in *Proc. ESM*, Warsaw, Poland, Jun. 1999, pp. 175–181.

[50] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research: Volume 1: Stochastic Processes and Operating Characteristics*. New York, NY, USA: Dover, 2003.

[51] 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.

[52] N. Hu, X.-L. Li, and Q.-N. Ren, "Random access preamble assignment algorithm of TD-LTE," in *Advan. Comput., Commun., Control Autom., Lec. Notes in Electr. Eng.*, 2012, vol. 121, pp. 701–708.

[53] M. Milosavljevic, M. Thakur, P. Kourtessis, J. Mitchell, and J. Senior, "Demonstration of wireless backhauling over long-reach PONs," *IEEE/OSA J. Lightwave Technol.*, vol. 30, no. 5, pp. 811–817, Mar. 2012.

[54] F. Aurzada, M. Scheutzow, M. Reisslein, N. Ghazisaidi, and M. Maier, "Capacity and delay analysis of next-generation passive optical networks (NG-PONs)," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1378–1388, May 2011.

[55] A. Bianco, T. Bonald, D. Cuda, and R.-M. Indre, "Cost, power consumption and performance evaluation of metro networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 5, no. 1, pp. 81–91, Jan. 2013.

[56] M. Scheutzow, M. Maier, M. Reisslein, and A. Wolisz, "Wavelength reuse for efficient packet-switched transport in an AWG-based metro WDM network," *IEEE/OSA J. Lightwave Technol.*, vol. 21, no. 6, pp. 1435–1455, Jun. 2003.

[57] H.-S. Yang, M. Maier, M. Reisslein, and W. M. Carlyle, "A genetic algorithm-based methodology for optimizing multiservice convergence in a metro WDM network," *IEEE/OSA J. Lightwave Technol.*, vol. 21, no. 5, pp. 1114–1133, May 2003.

[58] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

Revak R. Tyagi (S'09) received the B.E. degree in electronics and communications engineering from G.B. Pant Engineering College, Pauri-Garhwal, India, in 2005 and the M.S.E. degree from Arizona State University (ASU), Tempe, AZ, USA, in 2009. From 2005 to 2007, he was a Scientific Officer at Bhabha Atomic Research Center (BARC), Mumbai, India. He worked as a Graduate Research Assistant at The Biodesign Institute, ASU, during 2007–2010. He is currently working toward his Ph.D. and working as Graduate Research and Teaching Associate at ASU. His research interests are in the areas of radio access networks, wireless networking and mobile communications.

Sang G. Kim (S'92–M'00) is a principal researcher in LG Electronics Mobile Research (LGEMR) in San Diego, CA, USA. He received the M.S and D.Sc. degrees in electrical engineering from The George Washington University, Washington, DC, USA, and the B.S. and M.S. degrees in electronics engineering from Sung Kyun Kwan University, Seoul, Korea. He is actively involved in research and standard activities regarding a broad range of wireless communication systems. He is a chair of IEEE VTS San Diego Chapter. He was the recipient of the best paper award from the *Journal of Communications and Networks* in 1999.

Frank Aurzada studied mathematics at the University of Jena, Jena, Germany and the University of York, York, U.K. He received the Dipl.-Math. and Ph.D. degrees in mathematics from the University of Jena in 2003 and 2006, respectively. Then he was postdoctoral researcher at TU Berlin, Berlin, Germany, from 2006–2012 and Associate Professor at TU Braunschweig, Braunschweig, Germany, 2012–2013. Since 2013 he is Professor at the Technical University Darmstadt, Darmstadt, Germany.

Martin Reisslein (A'96–S'97–M'98–SM'03) is a Professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University (ASU), Tempe, AZ, USA. He received the Ph.D. in systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1998. His research interests are in the areas of multimedia networking, optical access networks, and engineering education.

Ki-Dong Lee (SM'07) received the B.S., M.S., and Ph.D. degrees from KAIST, Daejeon, Korea. His academy and industry experience includes design/analysis for satellite multimedia systems and 4G/5G cellular systems/service over the past 13 years in Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, University of British Columbia (UBC), Vancouver, BC, Canada, and LG Electronics, San Diego, CA, USA. His research interests include socio-economic approach to design/analysis of and value creation of brand-new mobile applications systems/services. He received awards from IEEE ComSoc, International Federation of Operational Research Scientists (IFORS)/Asia Pacific Operational Research Societies (APORS), SK Telecom, ETRI, and served as a co-guest editor for *IEEE Wireless Communications* and Technical Program Committee member for many IEEE-sponsored conferences.