Performance analysis of IEEE 802.11 WLANs with saturated and unsaturated sources

Suong H. Nguyen, Hai L. Vu, Lachlan L. H. Andrew
Centre for Advanced Internet Architectures, Technical Report 110811A
Swinburne University of Technology
Melbourne, Australia
hsnguyen@swin.edu.au, hvu@swin.edu.au, landrew@swin.edu.au

***Abstract*—This paper proposes a comprehensive but tractable model of IEEE 802.11 carrying traffic from a mixture of saturated and unsaturated (Poisson) sources, with potentially different QoS parameters, *TXOP limit*, $CW_{\min}$ and $CW_{\max}$. The model is used to investigate the interaction between these two types of sources, which is particularly useful for systems seeking to achieve load-independent "fair" service differentiation. We show that, when the *TXOP limit* for unsaturated sources is greater than one packet, batches are distributed as a geometric random variable clipped to *TXOP limit*. Furthermore, we present asymptotic results for the access delay distribution, which indicates that it is infeasible to obtain real-time service in the presence of 8 or more saturated sources regardless of the real time traffic load given that all stations use $CW_{\min}$ of 32.**

I. Introduction

In recent years, wireless local area networks (WLANs) have become very popular and are widely deployed, due to the rapid increase in demand for Internet access at any time and place through WiFi-enabled mobile devices such as laptops and personal digital assistants (PDAs). Internet applications over WLANs consist not only of throughput-intensive applications such as email, file transfer or web surfing but also of delay-sensitive ones such as voice and video. To provide quality of service (QoS) differentiation, IEEE 802.11e was specified in [1], which defines a contention-based medium access control (MAC) scheme called the Enhanced Distributed Channel Access (EDCA). In particular, EDCA provides service differentiation by the tuning of various MAC parameters: the minimum spacing between packets (Arbitration Inter-Frame Space or AIFS), minimum and maximum contention windows ($CW_{\min}$ and $CW_{\max}$), and lengths of packet bursts or transmission opportunity limit (*TXOP limit*).

In this paper, we model the performance of EDCA with a mixture of saturated non-realtime sources (i.e. each always has a packet to transmit) which seek high throughput, and unsaturated real-time sources which demand low delay. The motivation for our model is to enable the study of MAC mechanisms such as [18] that improve service for both types of users by means of three of the EDCA parameters: $CW_{\min}$ and $CW_{\max}$, which control how long a source waits before transmission, and *TXOP limit*, which controls how much it can transmit per channel access. We do not model AIFS because it provides load-dependent prioritization, which does not help to achieve the "fair" service differentiation we seek.

Before reviewing the related work, we first briefly describe the protocol and related concepts. Like the original Distributed Coordination Function (DCF) in IEEE 802.11, EDCA enables users to contend for the wireless channel using carrier sense multiple access with collision avoidance (CSMA/CA), with truncated binary exponential backoff (BEB) and slotted idle time. When a packet arrives to an idle source, it senses the channel for a period AIFS. If it is idle during this whole time, the packet is transmitted immediately (asynchronously). Otherwise, the source waits until the channel is continuously idle for AIFS, and then starts a backoff process. A backoff counter is initialized to a random integer uniformly distributed between 0 and $(CW-1)$, where CW is the current contention window. For each new packet, CW is initialized to $CW_{\min}$ and doubles after each unsuccessful transmission until it reaches $CW_{\max}$, after which it remains constant. The backoff counter is decreased by one at every idle slot time, of duration σ, and frozen during periods of channel activity. Decrementing is resumed one slot time before the expiration of an AIFS time after a channel activity period ends. (A subtle difference between EDCA and DCF is that in DCF, the decrementing is not resumed until after the expiration of AIFS [3].) When the backoff counter reaches zero, the source is allowed to transmit

for a *TXOP limit* period of time, which may allow one or more packets to be transmitted. (Note that the standard [1] also defines a mode in which only a single packet is transmitted when the backoff counter reaches zero.) An acknowledgment (ACK) is sent back from the receiver after a Short Inter-Frame Space (SIFS) for every successful packet reception. If an ACK is not received, the source increases CW as described above, and attempts again until the retry limit is reached. After receiving an ACK, the source performs a "post-backoff" process with contention window $CW_{\min}$ before being allowed to restart the above procedure. This prevents back-to-back packet transmission.

There has been much work in modeling DCF and EDCA using different approaches [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. It is usually assumed that if two sources transmit at the same time, they experience a collision, and both packets are lost. Most models are based on the approximation proposed in [2] that the collision probability of a transmission is constant and independent of the number of retransmissions suffered. To obtain the collision probability, a fixed point formulation was introduced in [2] describing the relationship between the collision probability and a so-called attempt probability. The latter is the probability that a source will attempt to send a packet in a slot, and is a function of both the collision probability and the evolution of the backoff process driven by its MAC (DCF or EDCA) parameters. Existing models can be classified by the traffic (saturated vs. unsaturated) and protocol issues (DCF vs. EDCA) they consider, and by whether they explicitly model backoff as a Markov chain or only require the mean value at each backoff stage (mean-based analysis). Our model is of the latter, simpler type but more comprehensive than existing models of that type. To clarify this contribution, we first look at what the current existing models offer.

Recall that we are interested in models of heterogenous users. Several models have been proposed for unsaturated traffic with heterogeneous arrival rates and packet sizes in single-class IEEE 802.11 DCF WLANs: [5] and [6] propose Markov chain models while [7] proposes a mean-based analysis. The former are derived from the saturated model in [2] by introducing to the Markov chain additional states representing an idle station. The latter uses the same approach of extending a saturated model, this time by introducing a conditional attempt probability conditioned on a source having a packet to send [21]. Conversely, saturated traffic can also be approximated by setting the probability a source has a packet to send at any given time to be 1 as suggested in [7].

Naturally, the above DCF models do not include *TXOP limit* and $CW_{\min}$ differentiation. Many existing models [8], [9], [10], [11], [12], [13], [14], [17], [15], [16] of IEEE 802.11e EDCA consider heterogeneous traffic differentiated by $CW_{\min}$ and AIFS; however, few explicitly consider *TXOP limit*. Among those that explicitly model *TXOP limit*, most such as [8], [10] are based on Markov chains. Few [13], [16] use mean-based analysis. It has been shown in [13] that creating an accurate model of *TXOP limit* differentiation requires more than simply inflating the packet length and is a nontrivial extension that requires careful consideration. We notice that in networks with large *TXOP limit*, two important aspects are missed in most previous models: the distribution of the number of packets sent per channel access (hereafter called "burst size") and the residual time of an ongoing transmission from other stations seen by a burst of an unsaturated source arriving during that transmission as a component of the burst's delay. The model in [10] captures the first aspect but requires a burdensome matrix calculation on each iteration when solving the fixed point. Besides, it does not capture the effect on the distribution of the loss probability due to exceeding retransmission limit.

The contributions of our paper are as follows

- We point out the need for modeling the residual time of an ongoing transmission when a burst of an unsaturated source arrives. We include that in our model.
- We propose a closed form expression for the distribution of the queue size of an unsaturated source, which allows us to obtain a closed form distribution of the burst size. Unlike prior work, we also capture the effect of loss probability on the distribution.
- We propose a simple method to approximate the distribution of access delay, which is much simpler than existing analysis using the z-transform [13], [17]. Based on the approximation, the slope of distribution curve's tail is easily obtained.
- We derive a lower bound on the number of saturated sources for which unsaturated sources of any load experience unacceptable delay.

The remainder of the paper is organized as follows. Section II introduces the notation and assumptions used in our model. Section III presents a new model of an 802.11e EDCA WLAN with both unsaturated and saturated sources. This model is validated in Section IV,

and applied in Section V to the analysis of the delay distribution.

II. Notation and modeling assumptions

We will model an 802.11 EDCA WLAN with $N_u \geq 0$ unsaturated Poisson sources (e.g. voice traffic) and $N_s \geq 1$ saturated sources (e.g. bulk data transfer), which by definition always have packets available for transmission. The model can easily be modified to describe DCF's different backoff rule.

The model assumes that channel conditions are ideal (no channel errors, hidden terminals or capture effect), so that packets are lost only due to collisions, which occur if and only if multiple sources transmit at the start of the same slot. In particular, if a burst from an unsaturated source arrives asynchronously (i.e., arrives at an empty queue and senses the channel idle for AIFS) then it is assumed that carrier sensing will prevent a collision, since other stations will not attempt to transmit until the next slot boundary, by which time they can sense this transmission. In our model, all sources use the basic access scheme without RTS/CTS and an AIFS that is equal to the Distributed Inter-Frame Space (DIFS). We also assume that all packets from a given source have equal size, and non-saturated sources can accommodate an arbitrary large number of packets (i.e., no buffer overflow).

Here we will use the same approximation proposed in [2] that the collision probability of a transmission is constant and independent of the number of retransmissions suffered. However, the model can also be modified to cover different collision probabilities at different transmission attempts of a burst as done in [22]; in this case, collision probability varies due to the Paradox of Residual Life [20].

The summary of notation used in our model is as follows

- Subscripts s and u, respectively, denote a generic saturated and unsaturated source; subscripts x and y both denote a generic source, which can be saturated or unsaturated.
- Superscripts i, c and s denote a quantity pertaining to a slot which is idle, a collision and a successful transmission, respectively.
- p_x and τ_x, respectively, are the collision probability and attempt probability of a source $x \in \mathbb{S} \cup \mathbb{U}$.
- L_u is the probability that the first packet of a burst from an unsaturated source $u \in \mathbb{U}$ is dropped due to exceeding retransmission limit.
- m and K, respectively, are the doubling limit and retransmission limit ($m \leq K$). The doubling limit is the maximum number of times a station doubles its contention window due to collision.
- Y is a random variable (r.v.) representing the duration of a generic slot, which is σ if the slot is idle, or longer if the slot is busy.
- Y_u is a r.v. representing a slot duration observed by a burst of the unsaturated source $u \in \mathbb{U}$ during its backoff.
- λ_u is the packet arrival rate of an unsaturated source $u \in \mathbb{U}$.
- W_x is the minimum contention window of the source $x \in \mathbb{S} \cup \mathbb{U}$.
- $U[a, b]$ denotes an integer uniformly distributed on $[a, b]$, $A \sim B$ denote that A and B are equal in distribution, and $\mathbb{E}[\cdot]$ denote the mean of a r.v..
- $U_{xj} \sim U[0, 2^{\min(j,m)} W_x - 1]$ $(0 \leq j \leq K)$ is a r.v. representing the number of backoff slots in the jth backoff stage of a burst from the source $x \in \mathbb{S} \cup \mathbb{U}$. In our model, U_{xj} is assumed to be independent of all random variables mentioned above.
- l_x is the size of a packet from the source $x \in \mathbb{S} \cup \mathbb{U}$.
- η_x is a r.v. denoting the number of packets per burst of the source $x \in \mathbb{S} \cup \mathbb{U}$ intending to transmit per channel access.
- r_x is the maximum number of packets which fit into *TXOP limit* of the source $x \in \mathbb{S} \cup \mathbb{U}$.
- ρ_u is the probability the non-saturated source $u \in \mathbb{U}$ has at least a burst in the queue at a given time.
- b_u is the probability that a burst arriving at the unsaturated source $u \in \mathbb{U}$, when the latter has no packets queued, finds the channel busy.
- T_x^s is the random time that a burst sent by a source $x \in \mathbb{S} \cup \mathbb{U}$ occupies the channel if it is successfully transmitted. It is related to physical 802.11 parameters by:

$$T_x^s = T_{\text{difs}} + \eta_x (T_{\text{p}x} + T_{\text{ack}}) + (2\eta_x - 1) T_{\text{sifs}} \quad (1)$$

  where T_{difs}, T_{sifs}, and T_{ack} are the duration of DIFS, SIFS, and transmission of an ACK packet, respectively, and $T_{\text{p}x}$ is the transmission time of a packet from the source x.
- T_x is T_x^s conditioned on $\eta_x = 1$, which is deterministic.

Note that the total collision time experienced by a source in a collision will be the transmission time of the longest packet involved in that collision, plus an Extended Inter-Frame Space (EIFS) for stations not involved in the collision or $ACKtimeout$ + DIFS for

stations involved in the collision. From [1], the duration of EIFS is

$$T_{\text{eifs}} = T_{\text{sifs}} + T_{\text{ack}} + T_{\text{difs}} \tag{2}$$

and the duration of $ACKtimeout$ can safely approximated by

$$T_{\text{ACKtimeout}} = T_{\text{sifs}} + T_{\text{ack}} \tag{3}$$

From above, it is clear that the total collision time experienced by a source in a collision is equal to T_x where x is the source with longest packet size among sources involved in the collision.

III. Model

The following model captures the interaction between unsaturated realtime traffic and saturated data flows in an 802.11e EDCA WLAN. Using the assumptions and notation described in Section II, the model takes the system parameters W_x, r_x, T_{px} ($x \in \mathbb{S} \cup \mathbb{U}$), and λ_u ($u \in \mathbb{U}$), as input, and predicts the throughput of the saturated sources and the access delay of the unsaturated sources.

Without loss of generality, define sources to be indexed in non-increasing order of packet size, regardless of whether they are saturated or unsaturated. That is, for all $x, y \in \mathbb{S} \cup \mathbb{U}$, $T_x \geq T_y$ for $x < y$.

A. Fixed point model

Central to the model is a set of fixed-point equations, where the collision probabilities of all sources are expressed in terms of the attempt probabilities of all sources, and vice versa. We will now derive the fixed point equations which will be presented in (14) below. Note that the collision probability of a source is the probability that source experiences a collision given that it is transmitting. Moreover, in our paper, the word "attempt" means the contention attempt, which is the attempt of the first packet of a burst.

First, to determine the collision probability, denote the probability that no sources transmit in a given slot by

$$G = \prod_{x \in \mathbb{S} \cup \mathbb{U}} (1 - \tau_x). \tag{4}$$

The collision probability of a given source $x \in \mathbb{S} \cup \mathbb{U}$ is then

$$p_x = 1 - \frac{G}{1 - \tau_x}. \tag{5}$$

Second, the attempt probability of a saturated source $s \in \mathbb{S}$ is the mean number of attempts per burst divided by the mean number of slots per burst

$$\tau_s = \frac{\sum_{k=0}^{K} p_s^k}{\sum_{k=0}^{K} (\mathbb{E}[U_{sk}] + 1) p_s^k} \tag{6}$$

where the mean number of backoff slots is

$$\mathbb{E}[U_{sk}] = \frac{2^{\min(k,m)} W_s - 1}{2}. \tag{7}$$

Next, we will determine the attempt probability of an unsaturated source. First note that the attempt probability of an arbitrary unsaturated source $u \in \mathbb{U}$ is the expected number of attempts per source u per second divided by the expected number of system slots per second, where the expected number of attempts per source u per second is the product of the expected number of bursts per source u per second and the expected number of attempts per burst. These are given as follows.

- The mean number of bursts per source u per second is its packet arrival rate λ_u divided by the mean size of a burst departing from the queue.
  The size of a burst departing from the queue is on average $\mathbb{E}[\eta_u]$ if the burst is successfully transmitted, or 1 if the retry limit is exceeded, since in that case only the head of line packet is dropped. Hence, its mean is

  $$L_u + (1 - L_u)\mathbb{E}[\eta_u] \tag{8}$$

  where

  $$L_u = p_u^{K+1} \tag{9}$$

  is the loss probability, and the mean size of a burst attempting to transmit in a given slot, $\mathbb{E}[\eta_u]$, depends on the queue size distribution at the node. For light load, $\mathbb{E}[\eta_u] = 1$; in general, it is given by (41) in Section III-C.
  Then, from (8), the mean number of bursts per source u per second is

  $$\frac{\lambda_u}{L_u + (1 - L_u)\mathbb{E}[\eta_u]} \tag{10}$$

- To determine the average number of attempts per burst from unsaturated sources, we make usual approximation [7], [10], [16], [19] that bursts from an unsaturated source arriving at an empty queue and sensing channel idle will contend for the channel, the same as when they arrive at non-empty queue or sense channel busy. Then, the mean number of attempts per burst from the source u is approximated by

  $$1 + \sum_{j=1}^{K} p_u^j = \frac{1 - p_u^{K+1}}{1 - p_u} \tag{11}$$

Simulation results suggest this is reasonably accurate, which appears to be due to the presence of saturated sources. This approximation is not required in the delay model of Section III-B.

- The mean number of system slots per second is

$$\frac{1}{E[Y]} \tag{12}$$

From (10), (11) and (12), the attempt probability of the source u is

$$\tau_u = \frac{\lambda_u}{L_u + (1-L_u)\mathbb{E}[\eta_u]} \frac{1-p_u^{K+1}}{1-p_u} \mathbb{E}[Y] \tag{13}$$

A special case of Eq. (13) in 802.11 DCF WLANs without saturated sources coincides with the model of [19].

The fixed point is between the collision probabilities in (5) and the attempt probabilities derived from (6) and (13):

$$\tau_s = 2(1-p_s^{K+1}) / \Big(W_s(1-(2p_s)^{m+1}) \frac{1-p_s}{1-2p_s} + (2^m W_s + 1)(1-p_s^{K+1}) - 2^m W_s (1-p_s^{m+1}) \Big), \quad s \in \mathbb{S} \tag{14a}$$

$$\tau_u = \frac{\lambda_u}{L_u + (1-L_u)\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1-p_u^{K+1}}{1-p_u}, \quad u \in \mathbb{U} \tag{14b}$$

$$p_x = 1 - \frac{G}{1-\tau_x}, \quad x \in \mathbb{S} \cup \mathbb{U}. \tag{14c}$$

The mean slot time $\mathbb{E}[Y]$ can be expressed in terms of the probability a^i that no sources transmit in a given slot, the probability a_x^s that a source x successfully transmits a burst in a given slot, and the probability a_x^c that there is a collision involving the source x and only sources $y > x$ with packets no larger than T_x. Specifically,

$$\mathbb{E}[Y] = a^i \sigma + \sum_{x \in \mathbb{S} \cup \mathbb{U}} a_x^s \mathbb{E}[T_x^s] + \sum_{x \in \mathbb{S} \cup \mathbb{U}} T_x a_x^c \tag{15a}$$

$$a^i = G \tag{15b}$$

$$a_x^s = \frac{\tau_x}{1-\tau_x} G \tag{15c}$$

$$a_x^c = \tau_x \prod_{y<x} (1-\tau_y) \left(1 - \prod_{y>x} (1-\tau_y) \right) = \frac{\tau_x}{1-\tau_x} \left(\prod_{y \le x} (1-\tau_y) - G \right) \tag{15d}$$

$$\mathbb{E}[T_x^s] = T_{\text{difs}} + \mathbb{E}[\eta_x](T_x + T_{\text{ack}}) + (2\mathbb{E}[\eta_x] - 1) T_{\text{sifs}} \tag{15e}$$

where G is given by (4). Note that all $N_s + N_u$ values of a_x^c can be calculated in $O(N_s + N_u)$ time, by the nested structure of the products in (15d).

The fixed point (14) involves $\mathbb{E}[\eta_x]$ and $\mathbb{E}[Y]$. For light load, $\mathbb{E}[\eta_x] = 1$; hence, solving the fixed point (14) requires only (15). In general, $\mathbb{E}[\eta_x]$ is given by (41) derived from the delay model; hence, solving the fixed point (14) requires (15) and the delay model in Sec. III-B.

Simpler form for $K = m = \infty$: Although the retransmission limit K is equal to 7 in 802.11 standards, in many settings a source rarely uses all 7 retransmissions. In that case, it is reasonable to reduce the complexity of the model by approximating K and m to be infinite. Then, the fixed point (14) simplifies to

$$\tau_s = \frac{2}{W_s \frac{1-p_s}{1-2p_s} + 1}, \quad s \in \mathbb{S} \tag{16a}$$

$$\tau_u = \frac{\lambda_u}{\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1}{1-p_u}, \quad u \in \mathbb{U} \tag{16b}$$

$$p_x = 1 - \frac{G}{1-\tau_x}, \quad x \in \mathbb{S} \cup \mathbb{U} \tag{16c}$$

B. Delay model

In this section, we calculate the access delay of bursts from an unsaturated source. This is not only an important performance metric for those sources, but is also used to determine $\mathbb{E}[\eta_x]$ used in the fixed point (14).

Define the access delay to be the time between the instant when the burst reaches the head of the queue and begins contending for the channel, and the time when it is successfully received. Note that our model assumes a packet is lost only due to exceeding the retry limit.

We first propose the access delay model of a burst that arrives at an empty queue. The novelty is that we capture two important features which cannot be ignored in that case: the behavior when the burst arrives at an idle channel, and the residual time of the busy period during which the burst arrives. The probability b_u that the burst arrives at a busy channel, and hence initiates a backoff process, can have an effect of up to 25% on the delay estimates when load is light. Hence, it is considered in our delay model, unlike in the fixed point model (14). Moreover, $T_{\text{res},u}$, a r.v. that represents the residual time of the busy period during which the burst arrived, is significant in the existence of sources with large *TXOP limit*. Prior work has neglected the effect of $T_{\text{res},u}$.

Let D_u be a r.v. representing the access delay of a burst from the unsaturated source $u \in \mathbb{U}$. Then

$$D_u = T_u^s + A_u. \tag{17}$$

Here the transmission time T_u^s, given by (1), is random since η_u is random. The r.v. A_u representing the total backoff and collision time of the burst before it is successfully transmitted has the distribution

$$A_u = \begin{cases} 0 & \text{w.p. } \dfrac{1-b_u}{1-b_u+b_u(1-p_u^{K+1})} \\ A_{uk} & \text{w.p. } \dfrac{b_u p_u^k(1-p_u)}{1-b_u+b_u(1-p_u^{K+1})} \end{cases}, \quad K \geq k \geq 0 \tag{18}$$

in which A_{uk} is a r.v. representing the total backoff and collision time of the burst provided that it is successfully transmitted in the kth backoff stage. The remainder of the complexity of the delay model comes from estimating the duration of the backoff slots which comprise A_{uk}.

Write

$$A_{uk} = \sum_{j=0}^{k} B_{uj} + \sum_{j=1}^{k} C_u + T_{\text{res},u} \tag{19}$$

where the r.v. B_{uj} accounts for the backoff time in the jth backoff stage; and the r.v. C_u represents the duration of a collision involving a tagged burst.

The backoff time B_{uj} is given by

$$B_{uj} = \sum_{k=1}^{U_{uj}} Y_{u,k} \tag{20}$$

where U_{uj} is the number of backoff slots in the jth backoff stage, and the $Y_{u,k} \sim Y_u$ are the independent, identically distributed (i.i.d.) durations of a slot conditional on source u not transmitting, namely

$$Y_u = \begin{cases} \sigma & \text{w.p. } a_u^i \\ T_x & \text{w.p. } a_{xu}^c, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \\ T_x^s & \text{w.p. } a_{xu}^s, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \end{cases} \tag{21}$$

where a_u^i, a_{xu}^c and a_{xu}^s are the probabilities, conditional on u not transmitting, of an idle slot, a collision between a source x and sources $y > x$ with packets no larger than T_x, and a success of a burst from a source x. a_u^i and a_{xu}^s are obtained by dividing the analogous quantities in (15b)–(15c) by $1-\tau_u$ while a_{xu}^c is given by

$$a_{xu}^c = \frac{\tau_x}{1-\tau_x}\left(\prod_{\substack{y \leq x \\ y \neq u}}(1-\tau_y) - \frac{G}{1-\tau_u}\right) \tag{22}$$

The random collision time C_u is the duration of the longest packet involved in a collision involving source u,

$$C_u = \max(T_u, T_x) \quad \text{w.p.} \quad a_{xu}^{cu}, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \tag{23}$$

where a_{xu}^{cu} is the probability that the source u collides with the source x and possibly sources $y > x$ with packets no larger than T_x, given by

$$a_{xu}^{cu} = \frac{\tau_x}{1-a_u^i} \prod_{\substack{y<x \\ y \neq u}} (1-\tau_y) \tag{24}$$

Finally, the probability b_u that the burst arrives during a busy slot can be estimated as

$$b_u = 1 - \frac{a_u^i \sigma}{\mathbb{E}[Y_u]} \tag{25}$$

Mean access delay: From (17), the mean access delay is

$$\mathbb{E}[D_u] = \mathbb{E}[A_u] + \mathbb{E}[T_u^s] \tag{26}$$

where by Wald's theorem for sums of i.i.d. random variables [23],

$$\begin{aligned} \mathbb{E}[A_u] \approx & \frac{b_u}{1-b_u+b_u(1-p_u^{K+1})}\Big(\frac{W_u}{2} \\ & .\Big(\frac{2(1-(2p_u)^{m+1})(1-p_u)}{1-2p_u} - 1 + p_u^{m+1} \\ & + (-1+2^{m+1}-m2^m)(p_u^{m+1}-p_u^{K+1}) + 2^m \\ & .\Big(\frac{p_u^{m+1}-p_u^{K+1}}{1-p_u} + mp_u^{m+1} - Kp_u^{K+1}\Big)\Big)\mathbb{E}[Y_u] \\ & + \mathbb{E}[C_u]\Big(\frac{1-p_u^K}{1-p_u}p_u - Kp_u^{K+1}\Big) \\ & + \mathbb{E}[T_{\text{res},u}](1-p_u^{K+1})\Big) \end{aligned} \tag{27}$$

The approximation in (27) comes from approximating

$$\mathbb{E}[U_{uj}] = \frac{2^{\min(j,m)}W_u - 1}{2} \approx 2^{\min(j,m)-1}W_u. \tag{28}$$

The mean slot duration $\mathbb{E}[Y_u]$ observed by the source u and the mean collision delay $\mathbb{E}[C_u]$ can be found from (21) and (23), respectively. The mean residual time $\mathbb{E}[T_{\text{res},u}]$ is given by [20]

$$\mathbb{E}[T_{\text{res},u}] = \frac{\mathbb{E}[Y_u^b]}{2} + \frac{\text{Var}[Y_u^b]}{2\mathbb{E}[Y_u^b]}, \tag{29}$$

where Y_u^b is the duration of a busy period caused by transmissions of other sources. Its distribution is similar to that of Y_u of (21), conditioned on the slot not being idle.

Simpler form for $K = m = \infty$: In this case, the mean access delay in (27) is reduced to

$$\mathbb{E}[A_u] \approx b_u \Bigg(\left(\frac{1}{2(1-2p_u)} \right) W_u \mathbb{E}[Y_u] + \frac{\mathbb{E}[Y_u]}{2(1-p_u)} + \frac{p_u}{1-p_u} \mathbb{E}[C_u] + \mathbb{E}[T_{\text{res},u}] \Bigg). \quad (30)$$

Remark 1: Although $\mathbb{E}[Y_u]$ and $\mathbb{E}[Y_u^b]$ can be calculated using (21), it is simpler to use

$$\mathbb{E}[Y_u] = \frac{\mathbb{E}[Y] - a_u^s \mathbb{E}[T_u^s] - \mathbb{E}[C_u]\tau_u p_u}{1-\tau_u}, \quad (31)$$

which comes from the fact that Y_u is Y excluding components involving the source u which are successful transmission of u or collision involving u and the fact that the probabilities a slot is idle, contains a successful transmission, or contains a collision among an arbitrary number of sources of Y_u are similar to those of Y scaled by $1-\tau_u$.

Then, $\mathbb{E}[Y_u^b]$ is given from $\mathbb{E}[Y_u]$ as

$$\mathbb{E}[Y_u^b] = \frac{\mathbb{E}[Y_u] - \sigma a_u^i}{1-a_u^i} \quad (32)$$

However, the form (21) is needed to calculate $\text{Var}[Y_u^b]$, and the distribution of delay as done in Appendix A.

Remark 2: Under high load, a burst of an unsaturated source is likely to see a non-empty queue when arriving. Hence, it will have queueing delay in addition to access delay.

In this case, the access delay model above can still be used, which can be justified as follows.

Under high load, there are three possibilities a burst from an unsaturated source will observe upon arriving:

- Empty queue and channel idle for DIFS. For this case, $A_u = 0$ but its probability is small. Hence, it is reasonable to approximate it by the first term of (18).
- Empty queue and channel idle less than DIFS. For this case, $A_u = A_{uk}$ with A_{uk} given in (19).
- Non-empty queue. For this case, $A_u = A_{uk}$ with A_{uk} given in (19) but without $\mathbb{E}[T_{\text{res},u}]$.

The last two cases can be reasonably approximated by the second term of (18) because A_u in these cases only differ in $\mathbb{E}[T_{\text{res},u}]$. Hence, the above delay model can be a good approximation under high traffic load. This will be confirmed by simulation in Sec. IV.

Note that the above delay model is not very accurate if $\mathbb{E}[T_{\text{res},u}]$ is significant compared with other components of the access delay, which is the case when the load from the tagged unsaturated source is high while the load from other stations in the network is light and other stations use very large *TXOP limit*.

To have more accurate calculation of the access delay under any load, the above delay model can be extended by modifying (18) and (19) as follows

- A_u in (18) now becomes A'_u given by

$$A'_u = \begin{cases} 0 & \text{w.p.} \quad (1-b_u)(1-\rho_u)/\Theta \\ A'_{uk} + \mathbb{E}[T_{\text{res},u}] & \text{w.p.} \quad b_u(1-\rho_u)/\Theta \\ A'_{uk} & \text{w.p.} \quad \rho_u p_u^k (1-p_u)/\Theta \end{cases} \quad (33)$$

  where $\Theta = (1-b_u)(1-\rho_u) + (1-(1-b_u)(1-\rho_u))(1-p_u^{K+1})$ and A'_{uk} $(0 \le k \le K)$ is given in (34) below.
- A_{uk} in (19) now becomes A'_{uk} given by

$$A'_{uk} = \sum_{j=0}^{k} B_{uj} + \sum_{j=1}^{k} C_u \quad (34)$$

The mean queueing delay can be straightforwardly calculated using the P-K formula for an M/G/1 queue with the mean and variance of the service time determined from the access delay model. However, that is out of scope of the present paper.

C. Distribution of burst size

1) Saturated sources: The burst size η_s of a saturate source $s \in \mathbb{S}$ is a constant and equal to r_s, the maximum number of packets that fit in *TXOP limit* of the source s. This is because a saturated source always has a packet waiting to transmit.

In particular, by (1),

$$\eta_s = r_s = \left\lfloor \frac{\text{TxOP limit} - T_{\text{difs}} + T_{\text{sifs}}}{T_{\text{px}} + T_{\text{ack}} + 2T_{\text{sifs}}} \right\rfloor \quad (35)$$

2) Non-saturated sources: A non-saturated source u will send in bursts up to r_u or the number of packets in the queue, whichever is less. To estimate the distribution of these burst sizes we first model the queue size process. Note that in this model, packets arrive separately. In practice, packets may arrive in bursts. The model could be extended to one such as [25], but that is out of the scope of this paper.

a) Distribution of queue size: Model the queue size process as a Markov chain, with state $k = 0, 1, 2...$ corresponding to having k packets in the queue. From state k, there are transitions at rate λ_u to state $k+1$ corresponding to packet arrivals. From state $k \ge 1$, there are transitions to state $k-1$ at rate $\mu_u L_u$, corresponding to the loss of a single packet due to excess collisions. In states $k = 1, ..., r_u$, all packets can form a single

batch, and so there are transitions to state 0 at rate $\mu_u(1-L_u)$, corresponding to the successful transmission of this batch. In states $k > r_u$, each batch consists of r_u packets and so there are transitions to state $k - r_u$ at rate $\mu_u(1 - L_u)$. This is illustrated in Fig. 1.

Note that this Markov approximation is only useful for estimating the queue distribution for low occupancies; we will show in Section V that the tail of the service time distribution can be heavy, which means this Markov approximation does not capture the tail properties of the queue size. However, the burst size distribution does not depend on the tail.

In the above Markov chain, the total service rate at each state is the same and determined by

$$\mu_k = \mu_u = \frac{1}{\mathbb{E}[D_u]}, \quad \forall k \geq 1 \tag{36}$$

where μ_k is the total service rate at state k; μ_u is the average service rate of an unsaturated source u; $\mathbb{E}[D_u]$ is its mean service time, which is mean access delay given by (26).

$$\mathbb{E}[D_u] = \mathbb{E}[A_u] + \mathbb{E}[T_u^s] \tag{37}$$

where $\mathbb{E}[T_u^s]$ is the mean successful transmission time of a burst, given by (1) with $\eta_u = \mathbb{E}[\eta_u]$.

As noted in [24], the service rate may actually differ between states. However, as will be shown by simulation below, the approximation of constant service rate is actually more accurate than the approximation in [24] under the considered circumstances, as well as being more tractable.

Let Q_u be a random variable representing the queue size of an unsaturated source u in this Markov model.

Observe that Fig. 1 is similar to that of bulk service systems presented in [20] where the service rate of all states are approximated to be equal to the average one, except the fact that there is a transition from every state k to the previous state $k-1$ which represents the case when the head of queue (HoQ) packet is dropped due to exceeding the retransmission limit. This suggests the following result.

Theorem 1: If $0 < \lambda_u < \mu_u(L_u + r_u(1 - L_u))$ then the above Markov chain has a geometric steady state distribution,

$$P[Q_u = k] = \Big(1 - \frac{1}{z_0}\Big)\Big(\frac{1}{z_0}\Big)^k, \quad k = 0, 1, 2, \cdots \tag{38}$$

where $z_0 > 1$ is a solution of

$$\rho_u z^{r_u+1} - (1 + \rho_u)z^{r_u} + L_u z^{r_u-1} + 1 - L_u = 0 \tag{39}$$

where $\rho_u = \lambda_u/\mu_u$.

Proof: The proof decomposes the transition matrix A of the Markov chain as the sum of those of an M/M/1 queue and a bulk service queue, with equal steady state distributions.

Let A'_x be the transition matrix of an M/M/1 queue with service rate $L_u\mu_u$ and arrival rate $x\lambda_u$, and A''_x be the transition matrix of a bulk service queue [20] with service rate $(1-L_u)\mu_u$ and arrival rate $(1-x)\lambda_u$. For $x \in (0, L_u\mu_u/\lambda_u)$, the M/M/1 queue has geometric steady state probabilities Q'_x whose mean q'_x increases continuously from 0 to ∞. For $x \in (1-(1-L_u)\mu_u/\lambda_u, 1)$, the bulk service queue has geometric steady state probabilities Q''_x whose mean q''_x decreases continuously from ∞ to 0. Let (a, b) be the intersection of those intervals. This is non-empty by the upper bound on λ_u. Then $q'_x - q''_x$ increases continuously on (a, b). It is negative as $x \to a$, as either $q'_a = 0$ if $a = 0$ or $q''_x \to \infty$ as $x \to \infty$ if $a > 0$. Similarly, it is positive as $x \to b$. Hence there is an $\tilde{x} \in (a, b) \subseteq (0, 1)$ such that $Q'_{\tilde{x}} = Q''_{\tilde{x}}$. Then $0 = Q'_{\tilde{x}}(A' + A'') = Q'_{\tilde{x}}A$, and so the geometric distribution $Q'_{\tilde{x}}$ is the steady state distribution of the original Markov chain.

Substitution of the form (38) into balance equations of the Markov chain, implies that z_0 is the solution greater than 1 of (39). ■

b) Distribution of burst size: Here we determine the distribution of burst size η_u that an unsaturated source u transmits whenever a burst is removed from the queue by successful transmission or the HoQ packet is removed from the queue due to exceeding the retry limit. This burst size is a function of the queue size. Since the transmission rate is equal (μ_u) in each state, the distribution of burst size η_u is equal to that of $\min(Q_u, r_u)$ conditioned on $Q_u \geq 1$, which has complementary cumulative distribution function (ccdf)

$$P[\eta_u > k] = \begin{cases} (1/z_0)^k & 0 \leq k < r_u \\ 0 & k \geq r_u. \end{cases} \tag{40}$$

Then, the mean burst size is the sum of its ccdf as follows.

$$\mathbb{E}[\eta_u] = \sum_{k=0}^{\infty} P[\eta_u > k] = \frac{1 - (1/z_0)^{r_u}}{1 - 1/z_0} \tag{41}$$

c) Comparison with other work: Note that [24] proposed a Markov chain of the queue size which is similar to the above except that it (a) assumes different service rates for different states, (b) ignores the transition when the retry limit is exceeded, and (c) has a finite buffer. Then, the distribution of queue size Q_u is determined by numerically solving the balance equations

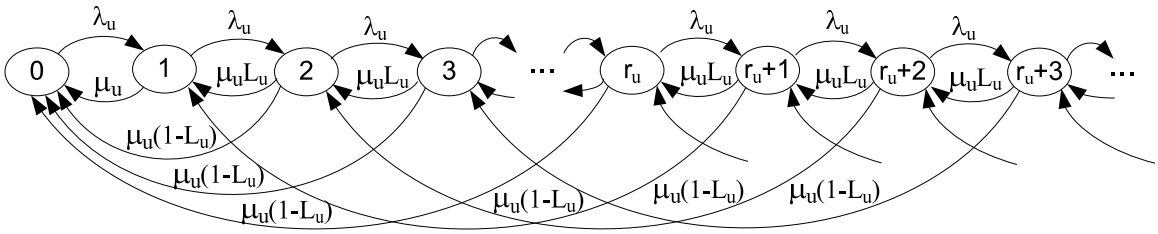


Fig. 1. The state-transition-rate diagram of queue size of an unsaturated source u.

and the distribution of burst size is approximated by the (time average) distribution of $\min(Q_u, r)$ conditioned on $Q_u > 0$.

One drawback of that approach is that it does not admit a closed-form solution for the distribution. Hence, it is computationally costly due to the matrix calculation on each iteration when solving the fixed point, especially when the buffer size of system is large. Since the model allows the service rates to differ in each state, it is natural to assume that the model of [24] would be more accurate (which would compensate for the higher complexity), but we now demonstrate that this is not the case.

Using the fixed-point model (14)–(15), we investigate the mean burst size $\mathbb{E}[\eta_u]$ determined from two Markov chains of queue size distribution: ours in Fig. 1 and the one in [24]. To have fair comparison, L_u is assumed to be 0 and the buffer capacity is set to be large (100 packets). The highest difference in $\mathbb{E}[\eta_u]$ between two Markov chains occurs when the network load is light and the arrival rate of a given unsaturated source is reasonably high. To investigate the accuracy of each scheme, we simulate such a scenario, specifically one with one saturated source and one unsaturated source with the arrival rate changing from small to large.

Note that [24] does not explicitly state how the service rate in each state is determined. Since it is constant for states greater than r_u, we assume that the service rate at state k satisfies

$$1/\mu_k = \mathbb{E}[A_u] + T_u^s|_{\eta_u=k}, \quad \forall k \geq 1 \tag{42}$$

where $T_u^s|_{\eta_u=k}$ is the duration of a successful transmission of a burst of k packets, given by (1) with $\eta_u = k$.

The results are in Fig. 2, which shows that $\mathbb{E}[\eta_u]$ from our Markov chain is closer to the simulation than that from the Markov chain of [24]. At this light load, the truncation to an occupancy of 100 packets is insignificant, and $L_u = 0$; hence, two Markov chains only differ in whether the service rate μ_k is constant or given by (42). This is counterintuitive, since (42) captures the increase in transmission time with k. We

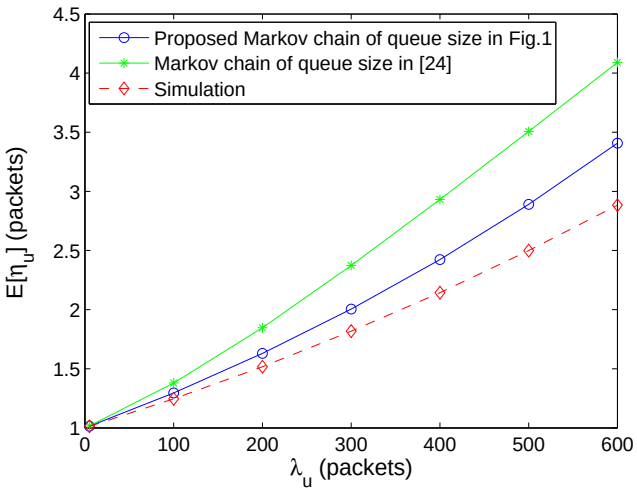


Fig. 2. The average burst size $\mathbb{E}[\eta_u]$ as a function of the arrival rate of an unsaturated source λ_u. (Unsaturated stations: Poisson arrivals with rate λ_u, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $r_u = 7$; Saturated stations: $N_s = 1$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

believe the inaccuracy is because (42) neglects the fact that the some fraction of the access delay $\mathbb{E}[A_u]$ has already elapsed by the time state k is reached, and so should not be reflected in (the reciprocal of) the transition rate. Since the true mean transmission time is the sum of an increasing term and a decreasing term, it is not clear *a prior* whether the constant rate μ_u or the increasing rate (42) would be a better model.

Another possible source of error is in obtaining the burst size distribution from the queue occupancy distribution. In [24] the burst size distribution was approximated by the *time average* distribution of $\min(Q_u, r)$ conditioned on $Q_u > 0$. However, the burst size depends on the queue size not at a typical point in time, but at a service instant. Thus, the weights given to different queue occupancies should be proportional to $\mu_k P[Q_u = k]$, rather than $P[Q_u = k]$. In our model, μ_k is independent of k and so these become equivalent.

D. Throughput of saturated sources

The throughput in packets/s of a saturated source $s \in \mathbb{S}$ is the average number of packets successfully transmitted per slot divided by the average slot length [2]

$$S_s = \frac{\mathbb{E}[\eta_s]\tau_s(1-p_s)}{\mathbb{E}[Y]} \tag{43}$$

where $\mathbb{E}[\eta_s]$ is the average number of packets per burst and the rest of the numerator is the probability the source s successfully transmits a burst in a given slot.

E. Model summary

The model described in the foregoing sections can be summarized as follows.

At low load, $\mathbb{E}[\eta_u] = 1$ for $u \in \mathbb{U}$; hence, the fixed point consists of (14), (15) and (35).

At high load, $\mathbb{E}[\eta_u]$ ($u \in \mathbb{U}$) depends on the distribution of queue size which involves the access delay; hence, the fixed point includes not only (14), (15) and (35) but also the delay model (17)–(29) and the burst size model (36)–(41).

The outputs p_x, τ_x ($x \in \mathbb{S} \cup \mathbb{U}$), S_s ($s \in \mathbb{S}$) and $\mathbb{E}[D_u]$ ($u \in \mathbb{U}$) can be determined by iteratively solving the fixed point numerically.

Consistency of the model: For our model to be physically meaningful, the rate of successful channel accesses per second of an unsaturated source should be less than that of a saturated source with the same $CW_{\min}$, m, and K.[1] When all sources have equal $CW_{\min}$, m, and K, this implies that for all $s \in \mathbb{S}$ and $u \in \mathbb{U}$,

$$\frac{\lambda_u}{\mathbb{E}[\eta_u]} < \frac{S_s}{\mathbb{E}[\eta_s]} \tag{44}$$

For situations where the burst arrival rate $\lambda_u/\mathbb{E}[\eta_u]$ does not satisfy this condition, an alternate instance of model (14)–(44) should be used, in which the unsaturated source u is replaced by a saturated source.

IV. Numerical Evaluation and Discussion

To validate the model consisting of (14)–(15),(17)–(29),(35)–(41), and (43), it was compared with simulations and, where possible, two existing models [5], [7]. The simulations used *ns-2.33* [29] with the EDCA package [30].

[1] It is not trivial that a saturated source achieves higher throughput than an unsaturated one; a network of only unsaturated sources can obtain a higher throughput than one of saturated sources [2, Fig. 3] because of the lower collision rate. However, within a given network, a saturated source gets a higher throughput than an unsaturated one with the same parameters.

TABLE I
MAC and PHYS parameters for 802.11b systems

Parameter	Symbol	Value
Data bit rate	r_{data}	11 Mbps
Control bit rate	r_{ctrl}	1 Mbps
PHYS header	T_{phys}	192 μs
MAC header	l_{mac}	288 bits
UDP/IP header	l_{udpip}	160 bits
ACK packet	l_{ack}	112 bits
Slot time	σ	20 μs
SIFS	T_{sifs}	10 μs
DIFS	T_{difs}	50 μs
Retry limit	K	7
Doubling limit	m	5
Buffer capacity		50 packets

We simulated networks of unsaturated sources and saturated sources sending packets to an access point using both DCF and EDCA. Under DCF, all stations are allowed to transmit only one packet per channel access; hence, $\eta_x = 1, \forall x \in \mathbb{S} \cup \mathbb{U}$. Recall that under EDCA, the standard [1] defines a mode which allows a source to transmit only one packet per burst, instead of specifying a duration *TXOP limit*. Throughout this section and Sec. V-B, we refer to this mode as $\eta = 1$.

All sources use the user datagram protocol (UDP). The traffic type used for unsaturated sources is either Poisson or quasi-periodic (CBR with some randomness in the inter-arrival time), as indicated in each scenario. Saturated sources receive CBR traffic at a rate faster than they can transmit. The MAC and physical layer parameters are set to the default values in IEEE 802.11b, as shown in Table I. These parameters determine T_x and T_x^s of (1) through the transmission duration of a packet of size l_x from a source x, and of an ACK packet,

$$T_{\text{px}} = T_{phys} + \frac{l_{mac} + l_{udpip} + l_x}{r_{data}}, \quad x \in \mathbb{S} \cup \mathbb{U}$$

$$T_{\text{ack}} = T_{phys} + \frac{l_{\text{ack}}}{r_{ctrl}}$$

In the figures of this section, simulation results are shown with confidence intervals which are determined using the Student t-distribution with confidence 95% [27]. Note that in some figures, the confidence intervals are too small to be seen.

A. Validation and comparison with existing DCF models

First, we compare our model with existing models of heterogeneous traffic [5], [7], which only consider IEEE 802.11 DCF, without multiple $CW_{\min}$ or *TXOP limit*s. To apply our model to DCF, we adjusted the backoff

decrement rule by replacing T_x^s and T_x ($x \in \mathbb{S} \cup \mathbb{U}$) in (15a) and (21) by $(T_x^s + \sigma)$ and $(T_x + \sigma)$.

1) Summary of two benchmark models: Before describing the simulation results, let us recall the models in [5] and [7].

a) Markov chain: The model in [5] is based on a Markov chain similar to that of [2], with additional states for unsaturated sources. It assumes that unsaturated sources have minimal buffers; therefore, when a packet arrives at a busy unsaturated source, it will be dropped. This causes the collision probability computed from this model to be smaller than that of models with non-zero buffers, such as our model. In [5] for heterogeneous traffic, the attempt probability of each type of traffic is obtained by solving a Markov chain and the collision probability of each type of traffic is determined as the probability that, when a source of that type transmits, there is at least one other source transmitting at the same time. The attempt and collision probabilities of each type of traffic are found by solving four simultaneous equations iteratively.

b) Mean-based: The model in [7] uses the mean-based approach for heterogeneous traffic, with an approximation in modelling an unsaturated source by a conditional attempt probability conditioned on the source having a packet to send. The approximation was first proposed in [21] for homogeneous traffic. This model assumes that unsaturated sources have infinite buffers.

The collision probabilities are again determined as the probability that, when a source of a given type transmits, there is at least one other source transmitting. Conversely, the attempt probability of each type of traffic is computed from the conditional attempt probability and the probability ρ that a source of that type has a packet to send at any given time. Because ρ depends on service time (which is the access delay in our model), this gives a fixed point involving two equations for collision probabilities and two equations for service times which can be solved iteratively. For a saturated source, $\rho = 1$.

It will be shown later in Figs. 3 and 4 below that the results of this model are not very accurate in the setting we consider. This appears to be a result of using the *time* average occupancy ρ instead of the probability that a source has a packet in a given *slot*. Thus we propose a modification to the model of [7] which replaces ρ by

$$\rho_{slot} = \frac{\lambda(\bar{w}_u + \mathbb{E}[R_u])}{S_s(\bar{w}_s + \mathbb{E}[R_s])} \tag{45}$$

where the numerator is the mean number of slots an unsaturated source has a packet per second and the denominator is the mean number of system slots per second. S_s and λ, respectively, are the throughput of a saturated source and the packet arrival rate of an unsaturated source; $\bar{w}_u$ and $\mathbb{E}[R_u]$, respectively, are the average total number of backoff slots and the average number of attempts a packet from an unsaturated source will encounter before it is successfully sent; and $\bar{w}_s$ and $\mathbb{E}[R_s]$ are the corresponding values for a saturated source. These quantities are already calculated in the model of [7]. Note that in this calculation, the packet service time of an unsaturated source is not used and as a result it need not be involved in the fixed point equations as it is in [7]. The proposed modification improves the match between the model of [7] and simulated values of the collision probabilities and throughput, but the match to mean access delay remains poor.

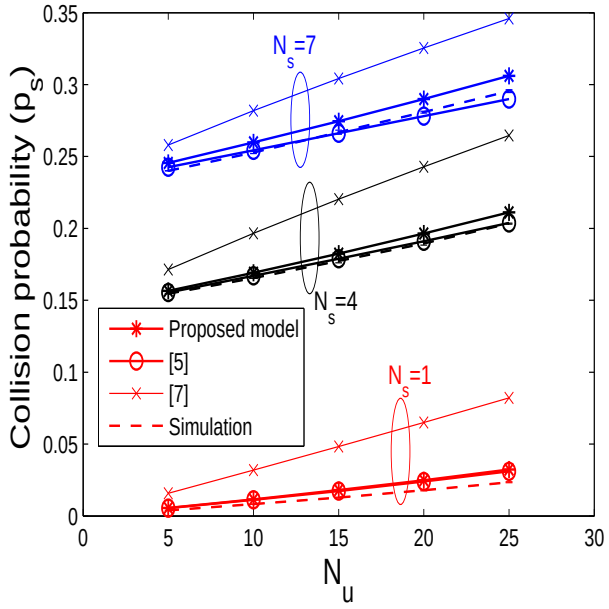
2) Validation: We simulated DCF networks of N_u identical unsaturated sources sending packets of size l_u with the Poisson arrival of rate λ, and N_s identical saturated sources sending packets of size l_s, all sending to an access point. We varied N_u, N_s, λ, l_u, and l_s. These two types of traffic have the same MAC parameters $\langle CW_{\min} = 32, \eta = 1 \rangle$ because there is no service differentiation in DCF.

Recall that subscripts s and u denote saturated and unsaturated sources, respectively.

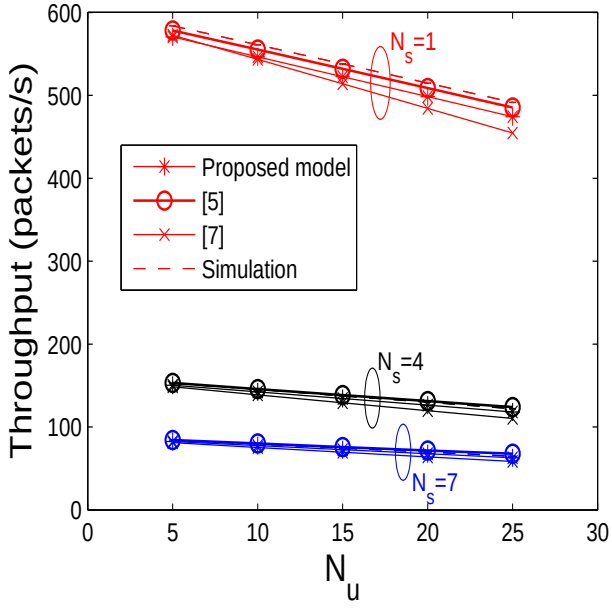
a) Scenario 1: In this scenario, we vary the number of saturated (N_s) and unsaturated (N_u) sources. The collision probability and throughput of a saturated source, and the collision probability and mean access delay of an unsaturated source, respectively, are shown in Fig. 3 as functions of N_u, parameterized by N_s. These figures show results computed from our model as well as from [5] and [7]. Simulation results are also plotted in the same figures.

Our model and the model of [5] accurately capture the increase in collision probabilities when N_s and N_u increases, and the resulting decrease in the throughput of saturated sources and increase in the mean access delay of unsaturated sources. However, the collision probabilities and mean access delay estimated from [7] are much higher than those of the simulation.

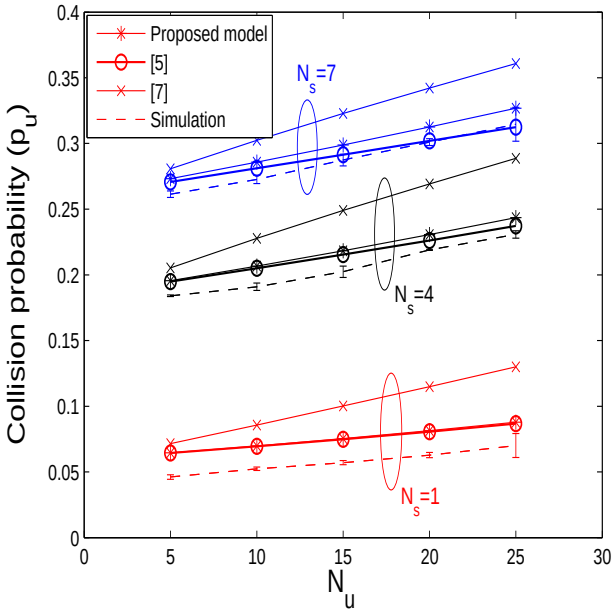
b) Scenario 2: In this scenario, we vary unsaturated sources' packet size (l_u, and hence T_u for $u \in \mathbb{U}$) and packet arrival rate (λ) while keeping the N_u and N_s unchanged. The collision probability and throughput of each saturated source, and the collision probability and mean access delay of an unsaturated source, respectively, are shown in Fig. 4 as functions of l_u, parameterized by λ. Results are obtained from our model, [5], [7], and



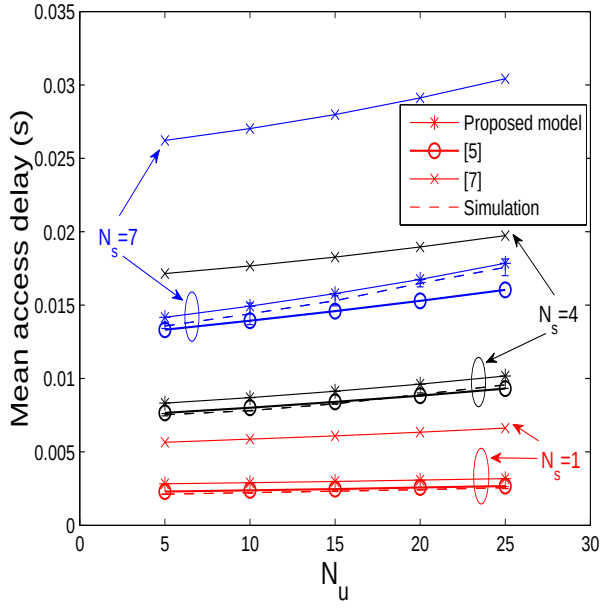
(a) Collision probability of a saturated source



(b) Throughput of a saturated source

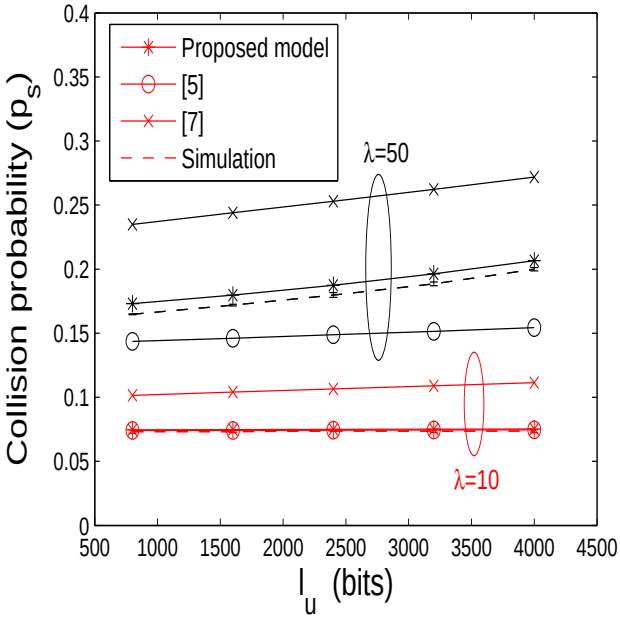


(c) Collision probability of an unsaturated source

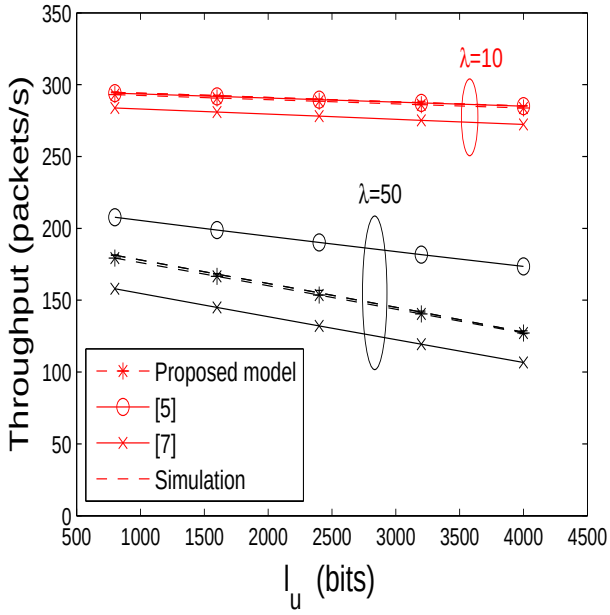


(d) Mean access delay of an unsaturated source

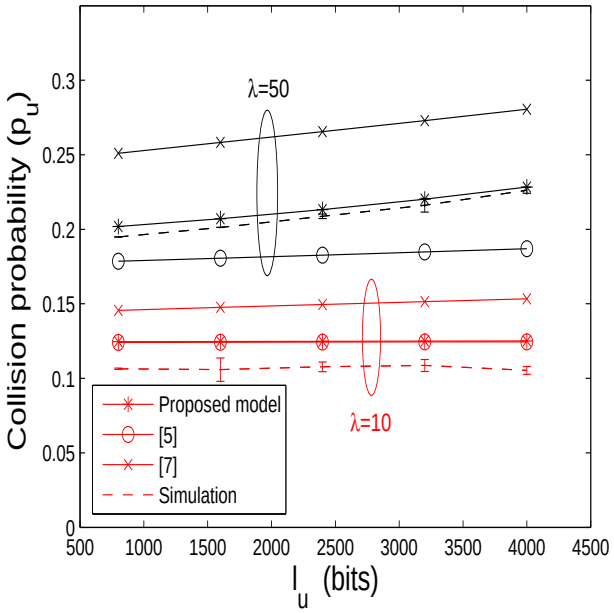
Fig. 3. Collision probabilities, throughput, and mean access delay for DCF, Scenario 1. Figs. 3(a), 3(c) and 3(d) clearly show that our model is much more accurate than the model in [7]. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)



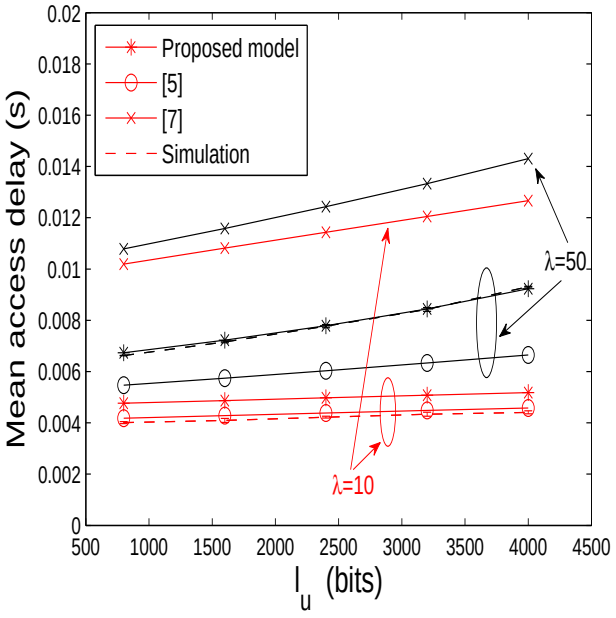
(a) Collision probability of a saturated source



(b) Throughput of a saturated source



(c) Collision probability of an unsaturated source



(d) Mean access delay of an unsaturated source

Fig. 4. Collision probabilities, throughput, and mean access delay for DCF, Scenario 2. Figs. 4(b) and 4(d), respectively, show clearly that our model is much more accurate than the models in [5] and [7]. (Unsaturated stations: Poisson arrivals with rate λ, $N_u = 10$, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 2$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

simulation.

Figure 4 shows that results from our model correctly capture the increase in collision probability with increasing l_u and λ, and the resulting decrease in throughput of saturated sources and increase in mean access delay. As for Scenario 1, the model in [7] overestimates the collision probabilities and mean access delay.

This scenario violates the zero-buffer assumption of [5], which hence becomes inaccurate when the packet arrival rate of unsaturated sources is 50 packets/s. That model predicts a high packet drop rate at high traffic load, which causes the collision probabilities to be underestimated.

In summary, our model for a network with both unsaturated and saturated sources developed in Section III is simple and versatile, and provides results more accurate than existing models when buffers are large.

B. Validation in 802.11e EDCA

In this subsection, we validate our model in 802.11e EDCA WLANs.

1) Scenario 3: We simulated networks with four types of traffic, denoted $u1$, $u2$, $s1$ and $s2$, of which the first two are unsaturated. The number of sources N, burst size η and packet size l will be distinguished by subscripts $u1$ to $s2$. Unsaturated sources of types $u1$ and $u2$ have different arrival rates λ_{u1} and λ_{u2}.

In this scenario, $N_{u1} = N_{u2} = N_{s1} = N_{s2} = N$, $l_{u1} = 500$ Bytes, $\lambda_{u1} = 10$ packets/s, $l_{u2} = 100$ Bytes, $\lambda_{u2} = 45$ packets/s, $l_{s1} = 1200$ Bytes, and $l_{s2} = 800$ Bytes. Packets arrive at unsaturated sources according to a Poisson process.

The QoS parameters $\langle CW_{\min}, \eta \rangle$ of sources of types $u1$, $u2$, $s1$ and $s2$, respectively, are $\langle 32, 2 \rangle$, $\langle 32, 5 \rangle$, $\langle 96, 1 \rangle$ and $\langle 96, 2 \rangle$.

The throughput in packets/s of a saturated source of type $s1$ and $s2$, and the mean access delay of an unsaturated source of type $u1$ and $u2$, respectively, are shown in Fig. 5 and Fig. 6 as a function of the number of sources per type N.

From Fig. 5, the throughput of a saturated source of type $s1$ is less than that of type $s2$. This is because types $s1$ and $s2$ have the same $CW_{\min}$ but type $s1$ has smaller *TXOP limit* and larger packet size. As can be seen, our model provides a surprisingly accurate estimate of the throughput.

Fig. 6 shows that our model also provides a reasonably accurate estimate of the mean access delay despite its simplicity compared with Markov chain based models.

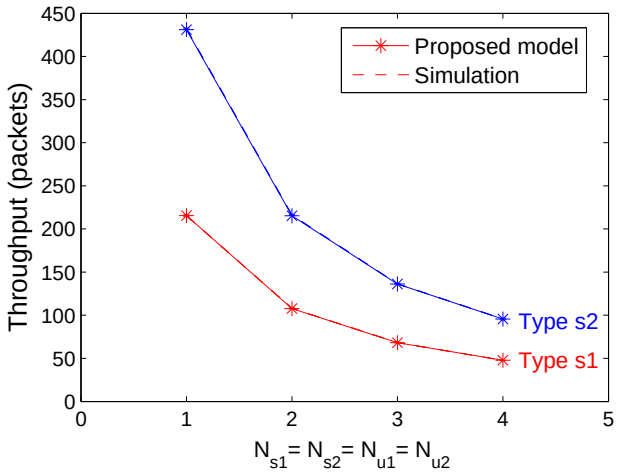


Fig. 5. Throughput of a saturated source of type $s1$ and $s2$, Scenario 3. (Unsaturated stations of type $u1$: Poisson arrivals with rate $\lambda_{u1} = 10$ packets/s, $l_{u1} = 500$ Bytes, $\eta_{u1} = 2$; Unsaturated stations of type $u2$: Poisson arrivals with rate $\lambda_{u2} = 45$ packets/s, $l_{u2} = 100$ Bytes, $\eta_{u2} = 5$; Saturated stations of type $s1$: $l_{s1} = 1200$ Bytes, $\eta_{s1} = 1$; Saturated stations of type $s2$: $l_{s2} = 800$ Bytes, $\eta_{s2} = 2$.)

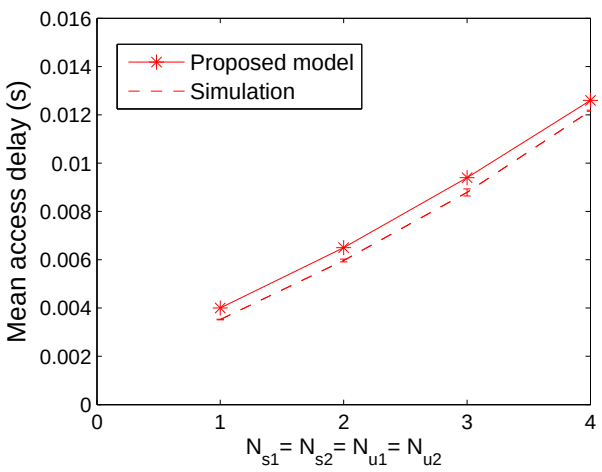
2) Scenario 4: We simulated networks of N_u identical unsaturated sources sending bursts of η_u packets of size l_u with the arrival rate λ, and N_s identical saturated sources sending fixed bursts of η_s packets of size l_s to an access point. Recall that subscripts s and u denote saturated and unsaturated sources, respectively.

Unsaturated sources have QoS parameters $\langle CW_{\min} = 32, \eta = 1 \rangle$. The QoS parameters of saturated sources are $\langle CW_{\min} = 32\eta_s, \eta = \eta_s \rangle$.

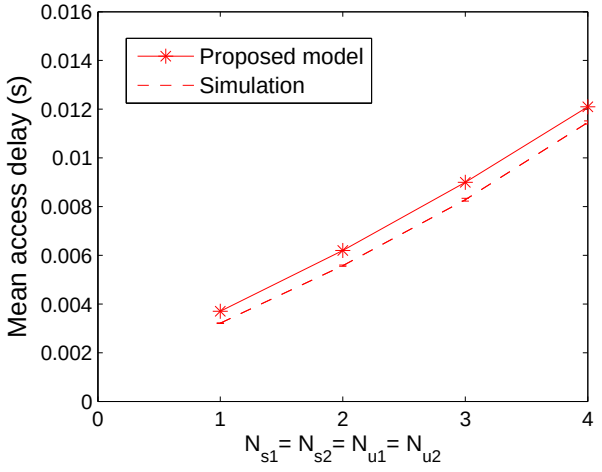
In this scenario, we vary the burst size of saturated sources (η_s). Also instead of Poisson arrivals, the packet inter-arrival times of unsaturated sources are set to be uniformly distributed in the range $1/\lambda \pm 1\%$. This quasi-periodic model represents voice traffic (which is often treated as periodic CBR traffic [31]), subject to jitter such as that caused by the operating system. Explicitly including this jitter is necessary to avoid "phase effect" artifacts in the results.

The throughput of a saturated source is shown in Fig. 7(a) as a function of η_s, parameterized by N_s. When η_s increases, there are fewer bursts from saturated sources contending for the channel, which decreases their collision probability. As a result, the throughput of a saturated source (in packets/s) increases.

One of the contributions of our proposed model is to capture the residual time of the busy period during which the burst arrived $T_{\text{res},u}$, which was not important in DCF, and has often been overlooked in EDCA models.

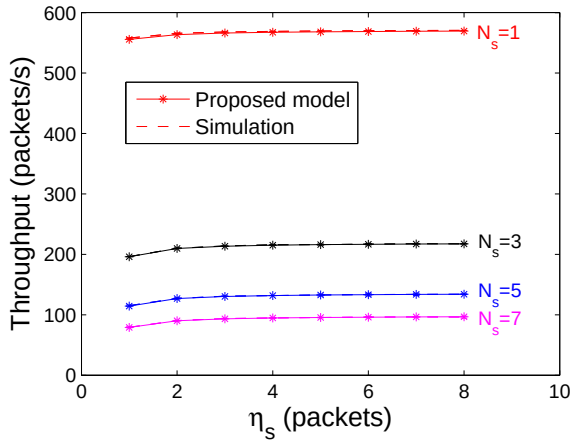


(a) Mean access delay of an unsaturated source of type $u1$

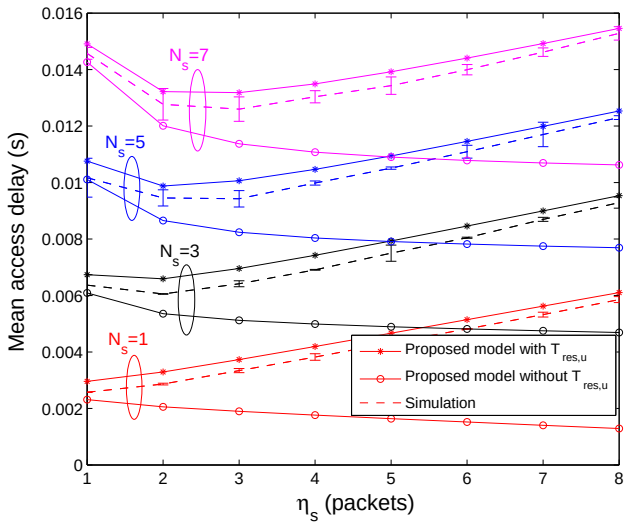


(b) Mean access delay of an unsaturated source of type $u2$

Fig. 6. Mean access delay of an unsaturated source of type $u1$ and $u2$, Scenario 3. (Unsaturated stations of type $u1$: Poisson arrivals with rate $\lambda_{u1} = 10$ packets/s, $l_{u1} = 500$ Bytes, $\eta_{u1} = 2$; Unsaturated stations of type $u2$: Poisson arrivals with rate $\lambda_{u2} = 45$ packets/s, $l_{u2} = 100$ Bytes, $\eta_{u2} = 5$; Saturated stations of type $s1$: $l_{s1} = 1200$ Bytes, $\eta_{s1} = 1$; Saturated stations of type $s2$: $l_{s2} = 800$ Bytes, $\eta_{s2} = 2$.)



(a) Throughput of a saturated source.



(b) Mean access delay of an unsaturated source

Fig. 7. Mean access delay and throughput when W_s and η_s are scaled together, Scenario 4. (Unsaturated stations: "quasi-periodic" traffic with rate $\lambda = 10$ packets/s, $N_u = 10$, $l_u = 200$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = \{1, 3, 5, 7\}$, $l_s = 1040$ Bytes, $W_s = \eta_s W_u$.)

Figure 7(b) shows the mean access delay of a burst from unsaturated sources with and without $T_{\text{res},u}$ in the access delay model under the same scenario. As can be seen, when η_s is large, $T_{\text{res},u}$ has significant effect on the delay estimation.

Also from Fig. 7(b), when η_s increases, for N_s greater than 1, there is a local minimum access delay. Initially, the dominant effect is the decrease in collisions due to the larger backoff window W_s of the saturated sources. For larger η_s, the increase in residual time $T_{\text{res},u}$ dominates this. This suggests that there is an optimal value for η_s where the access delay of unsaturated sources is minimum. This qualitative effect is not captured by models that neglect $T_{\text{res},u}$.

More importantly, Fig. 7 shows that increasing W_s and η_s together can benefit both unsaturated sources and saturated sources. Although the optimal value of η_s may vary in different scenarios, in most cases, η_s of 2 provides an improvement in the throughput of a saturated source and a reduction in mean access delay of unsaturated sources.

As our model is able to capture the right trend of mean access delay of unsaturated sources and throughput of saturated sources, it can be used to estimate the optimal η_s in this scenario.

V. Application of the Model

To demonstrate the usefulness of our model, we will use it to determine the distribution of access delay experienced by a burst from an unsaturated source. This is useful for tasks such as determining the appropriate size for jitter buffers.

For tractability, we approximate K and m to be infinite in the whole model and impose the approximation $b_u = 1$ in the delay model. Simulation results show that this gives accurate estimates of delay in the typical range of interest, from 10 ms to 1 s.

A. Analysis of access delay distribution

First note that the access delay distribution can be calculated using transform methods. The generating function of the complementary cumulative distribution function (ccdf) of access delay can be derived from its probability mass function (pmf). The distribution can then be obtained by numerical inversion of the z-transform, using, say, the Lattice-Poisson algorithm [26]. The details are not illuminating and are deferred to Appendix A. However, this demonstrates that this distribution information is embedded in our proposed model, unlike simpler models such as [7] which only consider the mean delay.

1) Approximation method: It is more informative to consider a simple approximate model of the access delay. The total burst access delay is the sum of many random variables: the backoff delays at each stage. However, at particular points, the ccdf of the access delay can be estimated accurately, from which the remainder can be estimated by interpolation. We will now derive such an approximation.

Let $W_{\text{med}}(k)$ be the median number of backoff slots used by bursts which succeed at the kth backoff stage (starting from $k = 0$). Since the number of slots at each stage j, U_{uj}, is symmetric about its median $M[U_{uj}] = (2^j W_u - 1)/2$, the median of their sum is

$$W_{\text{med}}(k) = \sum_{j=0}^{k} M[U_{uj}] = \left(2^k - \frac{1}{2}\right) W_u - \frac{k+1}{2}. \quad (46)$$

Note also that $W_{\text{med}}(k)$ is larger than $(2^k - 1)W_u - k$, the maximum number of backoff slots that could be experienced by a burst that succeeds at stage $k-1$ or earlier. It is possible for a burst which succeeds at stage $k+1$ or later also to experience $W_{\text{med}}(k)$ backoff slots but the probability of that is small, especially if p_u is small. Thus the unconditional ccdf of experiencing $W_{\text{med}}(k)$ backoff slots is slightly below the following upper bound

$$\begin{aligned} ccdf_W(W_{\text{med}}(k)) &\le 1 - \left(\sum_{j=0}^{k-1}(1-p_u)p_u^j + \frac{1}{2}(1-p_u)p_u^k\right) \\ &= p_u^k\left(\frac{1+p_u}{2}\right), \end{aligned} \quad (47)$$

which becomes tight for $p_u \ll 1$.

So far, this gives a good approximation for the ccdf of the number of backoff slots experienced. This can be related to the actual delay distribution by approximating the duration of each backoff slot by its mean, and adding the additional overhead of each stage. Thus, the delay associated with $W_{\text{med}}(k)$ backoff slots is approximately

$$\begin{aligned} D(W_{\text{med}}(k)) \approx& W_{\text{med}}(k)\mathbb{E}[Y_u] + k\mathbb{E}[C_u] + \mathbb{E}[T_{\text{res},u}] + \mathbb{E}[T_u^s] \\ =& 2^k W_u \mathbb{E}[Y_u] + k(\mathbb{E}[C_u] - \mathbb{E}[Y_u]/2) + K \\ \equiv& f(k) \end{aligned} \quad (48)$$

where Y_u is a slot duration observed by a burst of the unsaturated source u during its backoff and C_u is the duration of a collision involving a burst from the source u. The approximation becomes tight for large k by the law of large numbers.

This implies $k \approx f^{-1}(D(W_{\text{med}}(k))$, and so when $D = D(W_{\text{med}}(k))$ for some k,

$$ccdf_D(D) \approx \left(\frac{1+p_u}{2}\right) p_u^{f^{-1}(D)} \tag{49}$$

It turns out that (49) is a good approximation for any delay $D \geq D(W_{\text{med}}(0))$, rather than only the discrete points for which it was derived.

However, for delay $D < D(W_{\text{med}}(0))$, which corresponds to the total number of backoff slots from 0 to $W_u/2-1$, a much better approximation is possible. Note that the most likely way to back off for a small number of slots is to back off once, which gives a uniform distribution of the number of slots. Thus for $j = 0, 1, \ldots, W_u/2-1$, the ccdf of a delay

$$D(j) = j\mathbb{E}[Y_u] + \mathbb{E}[T_{\text{res},u}] + \mathbb{E}[T_u^s]$$

is approximately

$$\begin{aligned} ccdf_D(D(j)) \approx & 1 - (1-p_u)\frac{j+1}{W_u} \\ = & 1 - \frac{1-p_u}{W_u}\Big(1 \\ & + \frac{D(j) - \mathbb{E}[T_{\text{res},u}] - \mathbb{E}[T_u^s]}{\mathbb{E}[Y_u]}\Big). \end{aligned} \tag{50}$$

Thus, we propose the approximation that finds the ccdf from (50) for delays less than $D((W_u-1)/2)$, and from (49) for larger delays.

2) Power law delay distribution: In the proposed model, with unlimited retransmissions, the distribution of burst access delays has a power law tail ($At^kP(D > t) \to 1$ as $t \to \infty$ for some A, k). Although the true delay cannot be strictly heavy tailed when there is a finite limit on the number of retransmissions, the approximation holds for delays in the typical range of interest, from 10 ms to 1 s [32].

This power law arises since both the duration and probability of occurrence of the kth backoff stage increase geometrically in k. This is distinct from the heavy tailed delays occurring in ALOHA [33], which are caused by heavy-tailed numbers of identically distributed backoffs. Although the latter effect is very sensitive to the assumption of infinite retransmissions and the lack of burst fragmentation [34], 802.11 can be usefully modeled as heavy tailed even with typical limits of 6 to 8 retransmissions.

Note from (48) that $f(k) = 2^k W_u \mathbb{E}[Y_u] + O(k)$, where $h(m) = O(g(m))$ means that there exists a C such that for all sufficiently large m, $|h(m)| < Cg(m)$. Thus, by (49), the complementary CDF of a large delay D is approximately

$$ccdf_D(D) \approx \frac{1+p_u}{2}\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)^{\log_2(p_u)} \tag{51}$$

That is, the distribution has a power law tail with slope $\log_2(p_u)$, which increases (becomes heavier) with increasing congestion, as measured by the collision probability p_u. This is consistent with the more detailed calculations of [35]. Note that this insight would not be obtained by the direct use of the z-transform.

3) Excessive queueing delay: One application of the preceding result is to determine the congestion level at which the expected queueing delay for unsaturated sources becomes excessive. Although "excessive" will depend on the specific application, we will use the criterion that the expected queueing delay is infinite in our model with no limit on the binary exponential backoff. If each source is assumed to implement an M/G/1 queue, then this corresponds to the service time having infinite variance. (Note that the service time is the access delay in our model.)

Consider a log-log plot of the ccdf of a random variable D whose ccdf is the right hand side of (51). The minimum (steepest) slope for which the variance of D becomes infinite is -2 [35]. The right hand side of (51) suggests that this slope is $\log p_u / \log 2$. Thus the variance of D is infinite when $p_u \geq 2^{-2} = 1/4$. Under the model (16) and (43)–(44), we will now derive the minimum number of saturated sources N_s for which this occurs; that is, the N_s such that, for any number of unsaturated source N_u with arbitrary arrival rate, unsaturated sources which use the same backoff parameters as the saturated sources will have $p_u \geq 1/4$. Let us start with the following lemma, proved in Appendix B.

Lemma 1: Let s and u, respectively, denote an arbitrary saturated and unsaturated source. Under the model (16) and (43),

$$\frac{\tau_s}{\tau_u} = \frac{S_s\mathbb{E}[\eta_u]}{\lambda_u\mathbb{E}[\eta_s]}\frac{1-\tau_s}{1-\tau_u}.$$

If, in addition, (44) holds then $p_u > p_s$.

Theorem 2: Consider the model (16) and (43)–(44), with all sources using the same backoff parameters ($W_x = W, \forall x \in \mathbb{S} \cup \mathbb{U}$). If

$$N_s \geq 1 + \frac{\log(3/4)}{\log(1-\frac{4}{3W+2})} \tag{52}$$

then for any $N_u \geq 1$ and $\lambda_u > 0$, the variance of the random variable whose ccdf is the right hand side of (51) is infinite.

The proof is given in Appendix B. Surprisingly, the sufficient condition for infeasibility (52) depends only on W, the minimum contention window of all stations, and not settings such as the channel data rate, traffic of the real-time service, or the *TXOP limit*.

From (51), the distribution of an unsaturated source's access delay D_u under the model (16)–(44) has a tail which is approximately power law, given by the right hand side of (51). Hence, under the condition (52), the variance of the unsaturated source's access delay D_u is predicted to be infinite.

Note that the variance of the delays in the real system will not be infinite, due to the truncation of the backoff process. However, the high variability is enough to cause significant degradation of the user experience.

B. Numerical validation and discussion

This section has three objectives: (i) to validate the two methods of determining the distribution of access delay by comparing them with simulations; (ii) to validate the slope of the distribution curve's tail; (iii) to validate the condition (52) for the infinite variance of unsaturated sources' access delay.

The simulated network is the same as that in Sec. IV. Packets arrive at unsaturated sources according to a Poisson process. In the simulation of this section, all sources have the retry limit of 7 and the doubling limit of 5.

1) Validation of the distribution of access delay: The distribution of unsaturated sources' access delay determined from approximation and z-transform methods in comparison with simulation in different scenarios is shown in Figs. 8 and 9. Although derived assuming infinite retransmission, both the approximation and z-transform methods provide accurate estimates in the typical range of interest, from 10 ms to hundreds of ms. In particular, the approximation is of comparable accuracy to the z-transform method.

The big round markers on the approximation method curve show $D(W_{\text{med}}(k))$ of (48). In these scenarios, the approximation is quite accurate between the second and final attempts.

The approximation method inherits the limitations of our model on which it is based.

2) Slope of distribution curve's tail: The straight lines in Figs. 8 and 9 show the slope $\log_2(p_u)$. These capture the trend of the distribution curve in the typical range of interest, from tens to hundreds of ms.

3) Validation of Theorem 2: According to (52), when W is equal to 32 as in 802.11 DCF, the minimum number of saturated sources for the infinite variance of unsaturated sources' access delay is 8. This is validated in Fig. 10 which shows the distribution of access delay of unsaturated sources from NS-2 simulation. As can be seen, the slope of distribution curve's tail is slightly greater than -2 in the typical range of interest, from tens to hundreds of ms. This shows that these delays will occur as often as if the system had a power law tail with infinite variance.

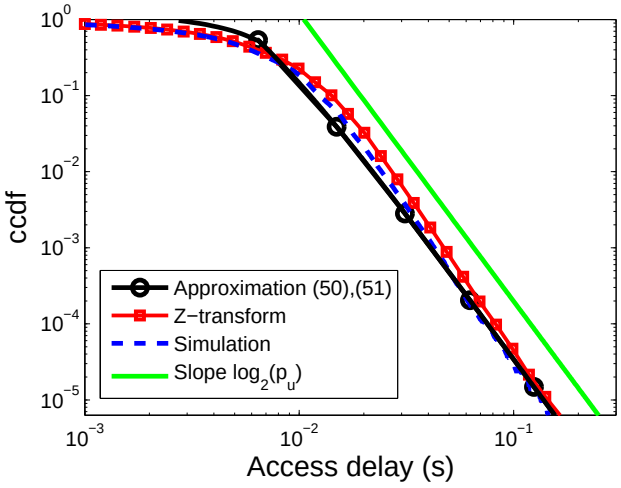


Fig. 8. Distribution of access delay. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $N_u = 15$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 2$, $l_s = 1040$ Bytes, $W_s = 3W_u$, $\eta_s = 4$.)

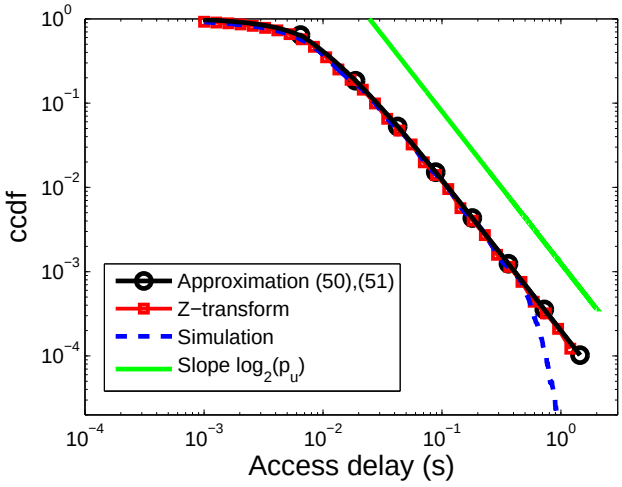


Fig. 9. Distribution of access delay. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $N_u = 20$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 6$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

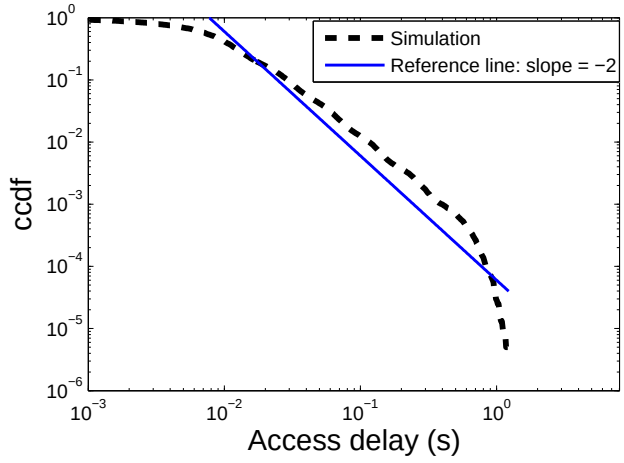


Fig. 10. Distribution of access delay of unsaturated sources. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 8$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

VI. Conclusion

We have provided a comprehensive but tractable fixed point model of 802.11 WLANs consisting of both unsaturated and saturated sources and shown that it provides accurate estimates of delay, throughput and collision probability in comparison with two existing models. We have proposed a closed form approximation for the distribution of the queue size of an unsaturated source, which is sufficiently accurate at low queue occupancies to predict the burst size distribution.

Using the fixed point model to investigate the interaction between these two types of traffic, we have briefly shown that "fair" service differentiation can be achieved based on two QoS parameters, *TXOP limit* and $CW_{\min}$. Moreover, a simple method to approximate access delay distribution has been proposed. Based on this approximation, the slope $\log_2(p_u)$ of distribution curve's tail has been obtained and then used to determine the lower bound on the number of saturated sources at which excessive queueing delay will be seen by unsaturated sources of arbitrary load, when all sources use the same MAC parameters.

Appendix A
The z-transform of delay

The generating function of the pmf of a non-negative integer-valued random variable X is given by

$$\hat{X}(z) = \sum_{k=0}^{\infty} P(X=k) z^k, \quad \text{for z} \in \mathbb{C} \tag{53}$$

To apply the z-transform, the continuous r.v.s D_u, A_u, A_{ui}, and $T_{\text{res},u}$ were quantized in steps of δ. Other random variables in Section III-B are non-negative and discrete, but some are not integer-valued. However, they can be transformed to integer-valued random variables, using the scale factor δ. Similarly, positive real variables such as σ, T_x, and T_x^s ($x \in \mathbb{S} \cup \mathbb{U}$) are also transformed to integers using δ.

By (17), the generating function of the access delay is

$$\hat{D}_u(z) = \hat{T}_u^s(z)\hat{A}_u(z) \tag{54}$$

Note that $\hat{T}_x^s(z)$ can be calculated from the distribution of η_x given by (40) for $x \in \mathbb{U}$ or (35) for $x \in \mathbb{S}$.

From (18), $\hat{A}_u(z)$ is given by

$$\hat{A}_u(z) = \frac{1}{1-b_u+b_u(1-p_u^{K+1})}\Big(b_u(1-p_u) \cdot \sum_{k=0}^{K} p_u^k \hat{A}_{uk}(z) + (1-b_u)\Big) \tag{55}$$

where, by (19), $\hat{A}_{uk}(z)$ can be approximated as

$$\hat{A}_{uk}(z) = \hat{C}_u(z)^k \hat{T}_{\text{res},u}(z) \prod_{j=0}^{k} \hat{B}_{uj}(z) \tag{56}$$

where $\hat{T}_{\text{res},u}(z)$ is [20]

$$\hat{T}_{\text{res},u}(z) = \frac{z}{(1-z)\mathbb{E}[Y_u^b]}\Big(1 - \frac{1}{1-a_u^i} \cdot \Big(\sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} a_{xu}^s \hat{T}_x^s(z) + \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} a_{xu}^c \hat{T}_x(z)\Big)\Big) \tag{57}$$

and

$$\hat{C}_u(z) = \frac{1}{1-a_u^i} \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} a_{xu}^{cu} \widehat{\max(T_u, T_x)}(z) \tag{58}$$

From (20), $\hat{B}_{uj}(z)$ is [28]

$$\hat{B}_{uj}(z) = \hat{U}_{uj}(\hat{Y}_u(z)) \tag{59}$$

where $\hat{U}_{uj}(z)$ is given by

$$\hat{U}_{uj}(z) = \frac{1 - z^{2^{\min(j,m)}W_u}}{2^{\min(j,m)}W_u(1-z)} \tag{60}$$

and $\hat{Y}_u(z)$ is determined from (21) as

$$\hat{Y}_u(z) = \hat{\sigma}(z)a_u^i + \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} a_{xu}^s \hat{T}_x^s(z) + \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} a_{xu}^c \hat{T}_x(z) \tag{61}$$

In summary, $\hat{D}_u(z)$ is given by

$$\hat{D}_u(z) = \Big(b_u(1-p_u)\sum_{k=0}^{K} p_u^k \hat{C}_u(z)^k \hat{T}_{\mathrm{res},u}(z)\prod_{j=0}^{k}\hat{U}_{uj}(\hat{Y}_u(z)) + (1-b_u)\Big)\frac{\hat{T}_u^s(z)}{1-b_u+b_u(1-p_u^{K+1})} \quad (62)$$

Then, the generating function of the ccdf, $\hat{D}_u^c(z)$ can be obtained from $\hat{D}_u(z)$ via the identity

$$\hat{D}_u^c(z) = \frac{1-\hat{D}_u(z)}{1-z}. \quad (63)$$

The access delay ccdf is the inverse z-transform of $\hat{D}_u^c(z)$.

Appendix B
Proof of Theorem 2

Proof of Lemma 1: Dividing p_s from (16c) by p_u from (16c), we have

$$\frac{1-p_u}{1-p_s} = \frac{1-\tau_s}{1-\tau_u} \quad (64)$$

Moreover, by (43),

$$\tau_s = \frac{S_s}{\mathbb{E}[\eta_s]}\frac{\mathbb{E}[Y]}{1-p_s} \quad (65)$$

Dividing (65) by τ_u from (16b), and applying (64) gives

$$\frac{\tau_s}{\tau_u} = \frac{S_s\mathbb{E}[\eta_u]}{\lambda_u\mathbb{E}[\eta_s]}\frac{1-p_u}{1-p_s} = \frac{S_s\mathbb{E}[\eta_u]}{\lambda_u\mathbb{E}[\eta_s]}\frac{1-\tau_s}{1-\tau_u} \quad (66)$$

which establishes the first claim.

By (44), this implies $\tau_s > \tau_u$, whence $p_u > p_s$ by (64). ■

Proof of Theorem 2: The result is a consequence of Lemma 1 and the following observations, which will be established below.

1) All else being equal, p_s is increasing in N_u.
2) If there are $N_u = 0$ unsaturated source and

$$N_s \geq 1 + \frac{\log(3/4)}{\log(1-\frac{4}{3W+2})} \quad (67)$$

then $p_s \geq 1/4$.
3) If $p_u > 1/4$ then the variance of the random variable whose ccdf is the right hand side of (51) is infinite.

These can be shown as follows:

1) This follows from (16c) since $\tau_u \in [0,1]$, and τ_s is decreasing in p_s.
2) When $N_u = 0$, (16c) becomes $p_s = 1-(1-\tau_s)^{N_s-1}$. Thus $p_s \geq 1/4$ if

$$\tau_s \geq 1 - \left(\frac{3}{4}\right)^{1/(N_s-1)}. \quad (68)$$

Conversely, (16a) is decreasing in p_s, and so $p_s \geq 1/4$ if

$$\tau_s \leq \frac{4}{3W+2} \quad (69)$$

Combining (68) and (69), $p_s \geq 1/4$ if

$$1 - \left(\frac{3}{4}\right)^{1/(N_s-1)} \leq \tau_s \leq \frac{4}{3W+2}$$

which upon rearrangement gives (67).
3) If $p_u > 1/4$, then the random variable whose ccdf is the right hand side of (51) has a tail heavier than kD_u^{-2} for some k, and hence infinite variance.

■

References

[1] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Standard 802.11e, 2005.

[2] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000.

[3] G. Bianchi, I. Tinnirello and L. Scalia, "Understanding 802.11e Contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations," *IEEE Network*, vol. 19, no. 4, pp. 28–34, 2005.

[4] N. Ramos, D. Panigrahi, and S. Dey, "Quality of Service Provisioning in 802.11e Networks: Challenges, Approaches, and Future Directions," *IEEE Network*, vol. 19, no. 4, pp. 14–20, 2005.

[5] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 Distributed Coordination Function in Nonsaturated Heterogeneous Conditions," *IEEE/ACM Trans. Networking*, vol. 15, no. 1, pp. 159–172, 2007.

[6] H. M. K. Alazemi, A. Margolis, J. Choi, R. Vijaykumar, and S. Roy, "Stochastic modelling and analysis of 802.11 DCF with heterogeneous non-saturated nodes," *Comp. Commun.,* vol. 30, pp. 3652-3661, 2007.

[7] X. Ling, L. X. Cai, J. W. Mark, and X. Shen, "Performance Analysis of IEEE 802.11 DCF with heterogeneous traffic," *Proc. IEEE Consumer Commun. Netw. Conf. (CCNC 2007)*, art. no. 4199105, pp. 49-53, 2007.

[8] I. Inan, F. Keceli, and E. Ayanoglu, "Modeling the 802.11e Enhanced Distributed Channel Access Function," *Proc. IEEE GLOBECOM'07*, pp. 2546–2551, 2007.

[9] B. Xiang, M. Yu-Ming, and X. Yun, "Performance Investigation of IEEE 802.11e EDCA under non-saturation condition based on the M/G/1/K model," *Proc. IEEE Conf. Industrial Electronics and Applications (ICIEA 2007)*, art. no. 4318419, pp. 298-304, 2007.

[10] J. Hu, G. Min, M. E. Woodward, and W. Jia, "A comprehensive analytical model for IEEE 802.11e QoS differentiation schemes under non-saturated traffic loads," *Proc. IEEE Int. Conf. Commun.*, art. no. 4533088, pp. 241-245, 2008.

[11] P. E. Engelstad and O. N. Østerbø, "Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction," *Proc. ACM MSWiM*, pp. 224-233, 2005.

[12] P. Serrano, A. Banchs, and A. Azcorra, "A Throughput and Delay Model for IEEE 802.11e EDCA under non-saturation," *Wireless Personal Commun.*, vol. 43, pp. 467–479, 2007.

[13] D. Xu, T. Sakurai, and H. L. Vu, "An Access Delay Model for IEEE 802.11e EDCA," *IEEE Trans. Mobile computing*, vol. 8, no. 2, pp. 261–275, 2009.

[14] J. Y. Lee and H. S. Lee, "A Performance Analysis Model for IEEE 802.11e EDCA Under Saturation Condition," *IEEE Trans. Commun.*, vol. 57, no. 1, 2009.

[15] J. Hui and M. Devetsikiotis, "A Unified Model for the Performance Analysis of IEEE 802.11e EDCA," *IEEE Trans. Communications*, vol. 53, no. 9, pp. 1498–1510, 2005.

[16] B. Bellalta, C. Cano, M. Oliver, and M. Meo, "Modeling the IEEE 802.11e EDCA for MAC parameter optimization," *3rd IEEE Consumer Communications and Networking Conference (CCNC) 2006*, vol. 1, pp. 390 - 394, 2006.

[17] I. Papapanagiotou, J.S. Vardakas, and G.S. Paschos, "Performance Evaluation of IEEE 802.11e based on ON-OFF Traffic Model," *Proc. ACM Int. Conf. Mobile multimedia communications*, 2007.

[18] S. H. Nguyen, H. L. Vu and L. L. H. Andrew, "Service Differentiation Without Prioritization in IEEE 802.11 WLANs," *Proc. IEEE LCN*, 2011.

[19] Q. Zhao, D. H. K. Tsang, and T. Sakurai, "A Simple and Approximate Model for Nonsaturated IEEE 802.11 DCF," *IEEE Trans. Mobile Computing*, vol. 8, no. 11, pp. 1539–1553, 2009.

[20] L. Kleinrock, "Queueing systems," John Wiley & Sons, Inc., NewYork, vol. 1, 1975.

[21] O. Tickoo and B. Sikdar, "Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC based Wireless Networks," *Proc. IEEE INFOCOM*, 2004.

[22] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew, "Packet size variability affects collisions and energy efficiency in WLANs," *Proc. IEEE Wireless Commun. Netw. Conf.*, 2010.

[23] S. M. Ross, *Introduction to Probability Models*, Academic Press, 2006.

[24] J. Hu, G. Min, and M. E. Woodward, "Analysis and Comparison of Burst Transmission Schemes in Unsaturated 802.11e WLANs," in *Proc. IEEE GLOBECOM*, pp. 5133-5137, 2007.

[25] W. B. Powell, "Iterative Algorithms for Bulk Arrival, Bulk Service Queues with Poisson and Non-Poisson Arrivals," *Transportation Science*, vol. 20, no. 2, pp. 65–79, 1986.

[26] Abbate J. and Whitt W., "Numerical inversion of probability generating functions," Operations Research Letters 12, pp. 245–251, 1992.

[27] W. A. Rosenkrantz, *Introduction to Probability and Statistics for Science, Engineering, and Finance*, CRC Press, 2008.

[28] O. C. Ibe, "Fundamentals of applied probability and random processes," Elsevier Inc., United Kingdom, 2005.

[29] "The network simulator ns-2," Available at http://www.isi.edu/nsnam/ns/.

[30] S. Wietholter and C. Hoene, "An IEEE 802.11e EDCF and CFB simulation model for ns-2," Available at http://www.tkn.tu-berlin.de/research/802.11e_ns2/.

[31] M. Menth, A. Binzenhfer, and S. Mhleck, "Source Models for Speech Traffic Revisited, " *IEEE/ACM Trans. Networking*, vol. 17, no. 4, pp. 1042–1051, 2009.

[32] J. Tan and N. B. Shroff, "Transition from heavy to light tails in retransmission durations," in Proc. IEEE INFOCOM, 2010.

[33] P. R. Jelenković and J. Tan, "Can retransmissions of superexponential documents cause subexponential delays?", *Proc. IEEE INFOCOM*, 2007.

[34] J. Nair, M. Andreasson, L. L. H. Andrew, S. H. Low and J. C. Doyle, "File Fragmentation over an Unreliable Channel", *Proc. IEEE INFOCOM*, 2010.

[35] J. Cho and Y. Jiang, "Basic theorems on the backoff process in 802.11," *ACM SIGMETRICS Perf. Eval. Review*, vol. 37, no. 2, pp. 18–20, 2009.