Performance analysis of IEEE 802.11 WLANs with saturated and unsaturated sources

Suong H. Nguyen, *Student Member, IEEE,* Hai L. Vu, *Senior Member, IEEE,*
and Lachlan L. H. Andrew, *Senior Member, IEEE*

***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.**

***Index Terms*—IEEE 802.11e EDCA, heterogeneous traffic.**

I. Introduction

Wireless local area networks (WLANs) are widely deployed to provide widespread Internet access through WiFi-enabled mobile devices such as laptops and smart phones. Internet applications over WLANs consist not only of throughput-intensive applications such as email and file transfer 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 Enhanced Distributed Channel Access (EDCA). EDCA allows service differentiation by tuning various MAC parameters: the minimum spacing between packets (Arbitration Inter-Frame Space or AIFS), the minimum and maximum contention windows ($CW_{\min}$ and $CW_{\max}$), and transmission opportunity limit (*TXOP limit*).

In this paper, we model 802.11 EDCA WLANs with a mixture of saturated non-realtime sources which seek high throughput, and unsaturated real-time sources which demand low delay. The motivation is to enable the study of MAC mechanisms such as [18] that improve service for both types of users by means of EDCA parameters: *TXOP limit*, $CW_{\min}$ and $CW_{\max}$. We do not model variable AIFS because it provides load-dependent prioritization, which does not help to achieve the "fair" service differentiation we seek.

A detailed description of the protocol and related concepts is presented in [1]. 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.

Existing models of DCF and EDCA [5–16] 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 recall existing models of heterogeneous 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 also extends a saturated model, this time by conditioning the attempt probability on a source having a packet to send [19]. Conversely, saturated traffic can 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 EDCA models [8–17] consider heterogeneous traffic differentiated by $CW_{\min}$ and AIFS; however, few explicitly consider *TXOP limit*. Among those that do, most such as [8, 10] are based on Markov chains. Few [13, 16] use mean-based analysis. Creating an accurate model of *TXOP limit* differentiation requires more than simply inflating the packet length [13]. Two important aspects of large *TXOP limit* are missed in most models: the distribution of the number of packets sent per channel access (the "burst size") and the residual time of an ongoing transmission from another station when a packet arrives at an idle station. The model in [10] captures the former but requires a burdensome matrix calculation on each iteration when solving the fixed point, and ignores the effect of loss on the distribution.

Our contributions are to: (1) model the residual time of an ongoing transmission in unsaturated sources' delay and show its importance; (2) calculate the distribution of the burst size of unsaturated sources; (3) propose a simple approximation to access delay distribution; (4) derive a lower bound on the number of saturated sources for which unsaturated sources experience unacceptable delay.

After introducing notation and assumptions in Section II, we present a model of EDCA WLANs with unsaturated and

The authors are with the Centre for Advanced Internet Architectures (CAIA), Faculty of Information and Communication Technologies, Swinburne University of Technology, Hawthorn, Melbourne VIC 3122, Australia. E-mail: {hsnguyen, hvu, landrew}@swin.edu.au.

saturated nodes in Section III, which is validated in Section IV. Delay asymptotics are studied in Section V.

II. Notation and modeling assumptions

We model an 802.11 EDCA WLAN with a set $\mathbb{U}$ of $N_u \geq 0$ unsaturated Poisson sources (e.g. voice traffic) and a set $\mathbb{S}$ of $N_s \geq 1$ saturated, bulk data sources, which always have packets to transmit.

The model assumes an ideal channel so that packets received correctly unless multiple sources transmit at the start of the same slot (a "collision"). Sources do not use RTS/CTS. All packets from a given source have equal size, and unsaturated sources can accommodate an arbitrary number of packets.

In the following description of notation, $s \in \mathbb{S}$, $u \in \mathbb{U}$ and $x, y \in \mathbb{S} \cup \mathbb{U}$ denote arbitrary sources, $U[a, b]$ denotes an integer uniformly distributed on $[a, b]$, $A \sim B$ denotes that A and B are equal in distribution, and $\mathbb{E}[\cdot]$ is ensemble average.

Source x emits packets of constant size l_x in bursts of a (possibly random) number of packets η_x, bounded above by the constant r_x.

The backoff mechanism imposes a slotted structure on time, with slot sizes independently distributed as a random variable Y, which is σ if the slot is idle or longer if a transmission is attempted. In each slot, x attempts to transmit with "attempt probability" τ_x and, conditional on making an attempt, collides with "collision probability" p_x. Following [2], these are assumed independent of the number of previous attempts of this packet, or packets from other stations.[1] If the first packet in the burst collides, the remainder are not transmitted. Transmissions of subsequent packets in a burst, not subject to contention, are not considered "attempts".

Each burst is attempted up to K times, with the jth attempt occurring after a backoff of $U_{xj} \sim U[0, 2^{\min(j,m)}W_x - 1]$ slots, where W_x is called the contention window. We assume U_{xj} is independent of random variables mentioned above. The size of a slot *conditioned* on source u performing a backoff is distributed as Y_u.

With probability L_x, all attempts of a burst suffer collisions, in which case the first packet is discarded.

Packets arrive to a source u as a Poisson process of rate λ_u and are queued. Source u has a packet to transmit a fraction ρ_u of the time. If a packet arrives when u has no packets to transmit, then with probability denoted $1 - b_u$ it observes the channel idle and transmits immediately. Such arrivals (termed "asynchronous") do not experience collisions, due to carrier sensing by the other stations at the start of the next slot.

Slots that are idle, collisions and successful transmissions are denoted by superscripts i, c, and s. The (random) time that a burst sent by a source x occupies the channel if it is successfully transmitted is given by

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

where T_{aifs}, T_{sifs}, and T_{ack} are the durations of AIFS, SIFS, and an ACK packet, and $T_{\text{p}x}$ is the transmission time of a packet from the source x. The deterministic value of T_x^s conditioned on $\eta_x = 1$ is denoted T_x.

The duration of a collision slot is the maximum of T_x over all sources x involved in the collision.[2]

III. Model

We now present a model that takes the system parameters W_x, r_x, $T_{\text{p}x}$, and λ_u, as input, and predicts the throughput of a source $s \in \mathbb{S}$ and the access delay of a source $u \in \mathbb{U}$.

Without loss of generality, sources are indexed in non-increasing order of packet size, regardless of whether they are saturated or unsaturated. That is, $T_x \geq T_y$ for $x < y$.

A. Fixed point model

The model is a set of fixed-point equations, where the collision probabilities are expressed in terms of the attempt probabilities, and vice versa. We will now derive the fixed point equations which will be presented in (9) below.

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

$$G = \prod\nolimits_{x \in \mathbb{S} \cup \mathbb{U}} (1 - \tau_x). \tag{2}$$

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

$$p_x = 1 - \frac{G}{1 - \tau_x}. \tag{3}$$

Second, the attempt probability of a saturated source s is the mean number of attempts per burst divided by the mean number of slots per burst

$$\tau_s = \frac{\sum_{k=0}^{K} p_s^k}{\sum_{k=0}^{K} (\mathbb{E}[U_{sk}] + 1) p_s^k} \tag{4}$$

where the mean number of backoff slots is

$$\mathbb{E}[U_{sk}] = 2^{\min(k,m)-1} W_s - 1/2. \tag{5}$$

Next, we determine τ_u, the attempt probability of an unsaturated source u. First, consider the number of packets u "serves" for each burst formed. With probability $L_u = p_u^{K+1}$, the first packet in the burst is discarded. Otherwise, u successfully sends on average $\mathbb{E}[\eta_u]$ packets. (The latter depends on the queue size distribution at the node; for light load, $\mathbb{E}[\eta_u] = 1$, and in general it is given by (31) in Section III-C.) Thus bursts are formed at rate

$$\frac{\lambda_u}{L_u + (1 - L_u)\mathbb{E}[\eta_u]}. \tag{6}$$

Next, determine the mean number of attempts per burst from u under the usual approximation [7, 10, 16, 22] that all bursts contend for the channel, even if they arrive asynchronously. The mean number of attempts is then

$$1 + \sum_{j=1}^{K} p_u^j = \frac{1 - p_u^{K+1}}{1 - p_u}. \tag{7}$$

[1] The model can be modified to reflect the fact that the residual life paradox [21] causes retransmissions to have different collision probabilities, as explained in [20].

[2] This is because stations involved in the collision wait for the ACK as usual, and other stations wait an Extended Inter-Frame Space (EIFS) [1].

Simulations 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.

From (6), (7) and the fact that there are $1/\mathbb{E}[Y]$ slots per second, the attempt probability of the source u is

$$\tau_u = \frac{\lambda_u}{L_u + (1-L_u)\mathbb{E}[\eta_u]} \frac{1-p_u^{K+1}}{1-p_u} \mathbb{E}[Y]. \tag{8}$$

A special case of Eq. (8) in 802.11 DCF WLANs without saturated sources coincides with the model of [22].

The fixed point is between the collision probabilities in (3) and the attempt probabilities derived from (4) and (8):

$$\tau_s = 2(1-p_s^{K+1}) / \Big(W_s(1-(2p_s)^{m+1}) \frac{1-p_s}{1-2p_s} + (2^m W_s + 1)(1-p_s^{K+1}) - 2^m W_s (1-p_s^{m+1}) \Big) \tag{9a}$$

$$\tau_u = \frac{\lambda_u}{L_u + (1-L_u)\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1-p_u^{K+1}}{1-p_u} \tag{9b}$$

$$p_x = 1 - \frac{G}{1-\tau_x}. \tag{9c}$$

The mean slot time $\mathbb{E}[Y]$ can be expressed in terms of the probabilities a^i, a_x^s and a_x^c that a given slot contains (a) no transmissions, (b) a successful burst transmission from source x, or (c) a collision involving the source x and only sources $y > x$ with packets no larger than T_x. Specifically,

$$\mathbb{E}[Y] = a^i \sigma + \sum_{x \in \mathbb{S} \cup \mathbb{U}} a_x^s \mathbb{E}[T_x^s] + \sum_{x \in \mathbb{S} \cup \mathbb{U}} T_x a_x^c \tag{10a}$$

$$a^i = G \tag{10b}$$

$$a_x^s = \frac{\tau_x}{1-\tau_x} G \tag{10c}$$

$$a_x^c = \frac{\tau_x}{1-\tau_x} \left(\prod_{y \le x} (1-\tau_y) - G \right) \tag{10d}$$

$$\mathbb{E}[T_x^s] = T_{\text{aifs}} + \mathbb{E}[\eta_x](T_x + T_{\text{ack}}) + (2\mathbb{E}[\eta_x] - 1) T_{\text{sifs}}. \tag{10e}$$

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 (10d).

The fixed point (9) involves $\mathbb{E}[\eta_x]$ and $\mathbb{E}[Y]$. For light load, $\mathbb{E}[\eta_x] = 1$; hence, solving (9) requires only (10). In general, $\mathbb{E}[\eta_x]$ is given by (31) derived from the delay model; hence, the delay model in Section III-B must be included.

Simpler form for $K = m = \infty$: Although the retry limit is $K = 7$ in 802.11, in many settings a source rarely uses all seven retransmissions. In that case, it is reasonable to reduce the complexity of the model by approximating K and m as infinite. Then, the fixed point (9) simplifies to

$$\tau_s = \frac{2}{W_s \frac{1-p_s}{1-2p_s} + 1}, \qquad s \in \mathbb{S} \tag{11a}$$

$$\tau_u = \frac{\lambda_u}{\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1}{1-p_u}, \qquad u \in \mathbb{U} \tag{11b}$$

$$p_x = 1 - \frac{G}{1-\tau_x}, \qquad x \in \mathbb{S} \cup \mathbb{U}. \tag{11c}$$

B. Delay model

We now calculate the access delay of bursts from an unsaturated source. This is not only an important performance metric, but also used to determine $\mathbb{E}[\eta_x]$ in (9). Access delay is defined to be duration 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.

We first propose an access delay model for a burst that arrives at an empty queue. The novelty is that we capture two important features in that case: the behavior when the burst arrives at idle channel, and the residual time of the busy period during which the burst arrives. The probability b_u that the burst arrives at busy channel can have an effect of up to 25% on the delay estimates when load is light. Moreover, the residual transmission time, $T_{\text{res},u}$, is significant in the presence of sources with large *TXOP limit*.

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

$$D_u = T_u^s + A_u \tag{12}$$

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

$$A_u = \begin{cases} 0 & \text{w.p. } \dfrac{1-b_u}{1-b_u+b_u(1-p_u^{K+1})} \\ A_{uk} & \text{w.p. } \dfrac{b_u p_u^k (1-p_u)}{1-b_u+b_u(1-p_u^{K+1})}, \ K \ge k \ge 0 \end{cases} \tag{13}$$

in which A_{uk} is the random 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

$$A_{uk} = \sum_{j=0}^{k} B_{uj} + \sum_{j=1}^{k} C_u + T_{\text{res},u} \tag{14}$$

where C_u is the random duration of a collision involving u, and the random the backoff time in the jth stage is

$$B_{uj} = \sum_{k=1}^{U_{uj}} Y_{u,k}. \tag{15}$$

Here 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

$$Y_u = \begin{cases} \sigma & \text{w.p. } a_u^i \\ T_x & \text{w.p. } a_{xu}^c, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \\ T_x^s & \text{w.p. } a_{xu}^s, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \end{cases} \tag{16}$$

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 (10b)–(10c) by $1-\tau_u$ while a_{xu}^c is given by

$$a_{xu}^c = \frac{\tau_x}{1-\tau_x} \left(\prod_{y \le x, y \ne u} (1-\tau_y) - \frac{G}{1-\tau_u} \right). \tag{17}$$

The random collision time C_u is the duration of the longest

packet involved in a collision involving source u,

$$C_u = \max(T_u, T_x) \quad \text{w.p.} \quad a_{xu}^{cu}, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \tag{18}$$

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

$$a_{xu}^{cu} = \frac{\tau_x}{1 - a_u^i} \prod_{\substack{y<x \\ y \neq u}} (1 - \tau_y). \tag{19}$$

Finally, the probability b_u can be estimated as

$$b_u = 1 - \frac{a_u^i \sigma}{\mathbb{E}[Y_u]}. \tag{20}$$

Mean access delay: From (12), the mean access delay is

$$\mathbb{E}[D_u] = \mathbb{E}[A_u] + \mathbb{E}[T_u^s]. \tag{21}$$

An explicit expression for $\mathbb{E}[A_u]$ is given in [23], using Wald's theorem [24] for (15). This is a function of the mean slot duration $\mathbb{E}[Y_u]$ seen by the source u, mean collision delay $\mathbb{E}[C_u]$ and mean residual time $\mathbb{E}[T_{\text{res},u}]$.

$\mathbb{E}[Y_u]$ and $\mathbb{E}[C_u]$ are found from (16) and (18), respectively. $\mathbb{E}[T_{\text{res},u}]$ is given by [21]

$$\mathbb{E}[T_{\text{res},u}] = \frac{\mathbb{E}[Y_u^b]}{2} + \frac{\text{Var}[Y_u^b]}{2\mathbb{E}[Y_u^b]}, \tag{22}$$

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 (16), conditioned on the slot not being idle.

Simpler form for $K = m = \infty$: The mean access delay again simplifies when K and m are infinite, becoming

$$\mathbb{E}[A_u] \approx b_u \Bigg(\left(\frac{1}{2(1-2p_u)} \right) W_u \mathbb{E}[Y_u] + \frac{\mathbb{E}[Y_u]}{2(1-p_u)} + \frac{p_u}{1-p_u} \mathbb{E}[C_u] + \mathbb{E}[T_{\text{res},u}] \Bigg). \tag{23}$$

Remark 1: Although $\mathbb{E}[Y_u]$ and $\mathbb{E}[Y_u^b]$ can be calculated using (16), it is simpler to use

$$\mathbb{E}[Y_u] = \frac{\mathbb{E}[Y] - a_u^s \mathbb{E}[T_u^s] - \mathbb{E}[C_u]\tau_u p_u}{1 - \tau_u}, \tag{24}$$

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

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

However, the form (16) is needed to calculate $\text{Var}[Y_u^b]$, and the distribution of delay as done in [23].

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. 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.

To see that the access delay model above can still be used in the presence of queueing, note that there are three possibilities a packet arriving to an unsaturated source can observe:

- Empty queue and channel idle for AIFS. For this case, $A_u = 0$ as in the first case of (13).
- Empty queue but channel not idle for AIFS. For this case, $A_u = A_{uk}$ with A_{uk} given in (14).
- Non-empty queue. For this case, $A_u = A_{uk}$ with A_{uk} given in (14) but without $\mathbb{E}[T_{\text{res},u}]$.

The last two cases can be approximated by the second term of (13) when $\mathbb{E}[T_{\text{res},u}]$ is small. The probability of $A_u = 0$ is slightly over-estimated by (13), but this effect is small at high load, since $b_u \to 1$ as load increases. It is confirmed by simulation in Section IV that (13) is often a good approximation for delay at high load.

Note that the above delay model becomes inaccurate in the uncommon case that $\mathbb{E}[T_{\text{res},u}]$ is significant compared with the access delay, which occurs when the arrival rate from source u is high while the arrival rate from other stations is light and other stations use very large *TXOP limit*. A more accurate by less tractable model is obtained by replacing (14) and (13) by

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

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

where $\Theta = (1-b_u)(1-\rho_u) + (1-(1-b_u)(1-\rho_u))(1-p_u^{K+1})$.

C. Distribution of burst size

1) Saturated sources: The burst size η_s of a saturate source 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),

$$\eta_s = r_s = \left\lfloor \frac{\text{TxOP limit} - T_{\text{aifs}} + T_{\text{sifs}}}{T_{\text{px}} + T_{\text{ack}} + 2T_{\text{sifs}}} \right\rfloor. \tag{26}$$

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 the Markov chain in Fig. 1, with state $k = 0, 1, 2 \ldots$ 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 $k \geq 1$, 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 $k = 1, \ldots, r_u$, all packets can form a single batch, and so there are transitions to state 0

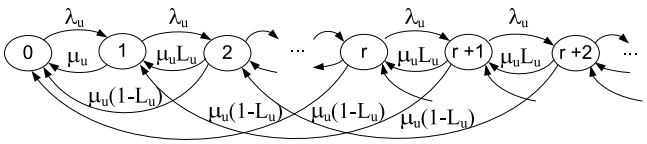


Fig. 1. The transition diagram of queue size of an unsaturated source u.

at rate $\mu_u(1-L_u)$ due 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 $\mu_u(1-L_u)$. 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

$$\mu_k = \mu_u = 1/\mathbb{E}[D_u], \quad \forall k \geq 1 \tag{27}$$

where μ_k is the total service rate at state k; μ_u is the mean service rate of source u; $\mathbb{E}[D_u]$ is given by (21).

As noted in [26], 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 [26] 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 in [21], except there is an additional transition from every state k to the previous state $k-1$ which represents the case when the head of queue packet is dropped due to exceeding retry 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,

$$P[Q_u = k] = \Big(1 - \frac{1}{z_0}\Big)\Big(\frac{1}{z_0}\Big)^k, \quad k = 0, 1, 2, \cdots \tag{28}$$

where $z_0 > 1$ is a solution of

$$\rho_u z^{r_u+1} - (1+\rho_u)z^{r_u} + L_u z^{r_u-1} + 1 - L_u = 0 \tag{29}$$

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 [21] 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 (28) into balance equations of the Markov chain, implies that z_0 is the solution greater than 1 of (29). ■

b) Distribution of burst size: Here we determine the distribution of burst size η_u of an unsaturated source u, which 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)

$$P[\eta_u > k] = \begin{cases} (1/z_0)^k & 0 \leq k < r_u \\ 0 & k \geq r_u. \end{cases} \tag{30}$$

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

$$\mathbb{E}[\eta_u] = \sum_{k=0}^{\infty} P[\eta_u > k] = \frac{1-(1/z_0)^{r_u}}{1-1/z_0} \tag{31}$$

c) Comparison with other work: [26] proposed a Markov chain of the queue size 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 balance equations 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 matrix calculation on each iteration when solving the fixed point, especially when the buffer size is large.

Using the fixed-point model (9)–(10), 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 [26]. 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 source u is reasonably high. We simulate such a scenario, specifically one with one saturated source and one unsaturated source with the arrival rate changing from small to large.

It is not explicitly stated in [26] 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

$$1/\mu_k = \mathbb{E}[A_u] + T_u^s|_{\eta_u=k}, \quad \forall k \geq 1 \tag{32}$$

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 in Fig. 2 shows that $\mathbb{E}[\eta_u]$ from our Markov chain is closer to the simulation than that from the Markov chain of [26]. At this light load, the truncation to an occupancy of 100 packets is insignificant, and $L_u = 0$; hence, the two Markov chains only differ in whether the service rate μ_k is constant or given by (32). We believe the inaccuracy of [26] is because (32) neglects the fact that some fraction of the access delay $\mathbb{E}[A_u]$ has already elapsed by the time state k is

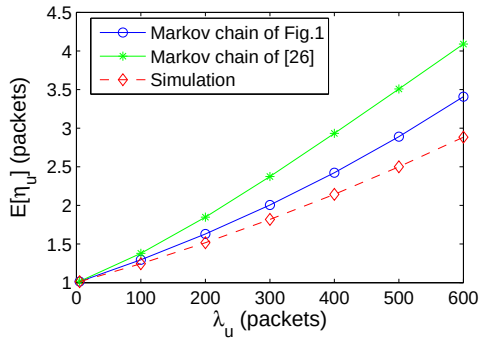


Fig. 2. The average burst size $\mathbb{E}[\eta_u]$ as a function of the arrival rate of an unsaturated source λ_u. (Unsaturated stations: Poisson arrivals with rate λ_u, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $r_u = 7$; Saturated stations: $N_s = 1$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

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 (32) would be a better model.

Another possible source of error is in obtaining the burst size distribution from the queue occupancy distribution. In [26] 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]

$$S_s = \frac{\mathbb{E}[\eta_s]\tau_s(1-p_s)}{\mathbb{E}[Y]}. \tag{33}$$

E. Model summary

Our model from previous sections is summarized as follows.

At low load, $\mathbb{E}[\eta_u] = 1$ for $u \in \mathbb{U}$; hence, the fixed point consists of (9), (10) and (26).

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 (9), (10) and (26) but also the delay model (12)–(22) and the burst size model (27)–(31).

The outputs p_x, τ_x, S_s and $\mathbb{E}[D_u]$ can be determined by iteratively solving the fixed point numerically and applying (33).

Consistency of the model: For our model to be physically meaningful, the rate of successful channel accesses per second of source u should be less than that of a saturated source with the same $CW_{\min}$, m, and K.[3] When all sources have equal $CW_{\min}$, m, and K, this implies that for all $s \in \mathbb{S}$ and $u \in \mathbb{U}$,

$$\frac{\lambda_u}{\mathbb{E}[\eta_u]} < \frac{S_s}{\mathbb{E}[\eta_s]}. \tag{34}$$

For situations where the burst arrival rate $\lambda_u/\mathbb{E}[\eta_u]$ does not satisfy (34), an alternate instance of model (9)–(34) should be used, in which source u is replaced by a saturated source.

[3]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
AIFS	T_{aifs}	50 μs
Retry limit	K	7
Doubling limit	m	5
Buffer capacity		50 packets

IV. Numerical Evaluation and Discussion

To validate the model (9)–(10),(12)–(22),(26)–(31), and (33), it was compared with simulations (using *ns-2.33* [30] and [31]) and, where possible, two existing models [5], [7].

We simulated networks of unsaturated and saturated sources sending packets to an access point using DCF and EDCA. All sources use the user datagram protocol (UDP). Unsaturated sources use either Poisson or quasi-periodic traffic (CBR with randomness in inter-arrival time). Saturated sources receive CBR traffic faster than they can transmit. We use the 802.11b parameters in Table I. The T_x and T_x^s in (1) are

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

Simulation results are shown with 95% Student-t confidence intervals [28]. In some figures, the confidence intervals are too small to be seen.

A. Validation and comparison with existing DCF models

Here our model is compared with existing models for heterogeneous traffic [5, 7] using 802.11 DCF. To apply our model to DCF, we adjusted the backoff decrement rule by replacing T_x^s and T_x in (10a) and (16) by $(T_x^s+\sigma)$ and $(T_x+\sigma)$.

1) Summary of two benchmark models: We first 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 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.

b) Mean-based: In [7] the mean-based approach is used for heterogeneous traffic where the attempt probability of an unsaturated source is multiplied by the probability ρ that the source having a packet to send. For saturated sources, $\rho = 1$. Unsaturated sources are assumed to have infinite buffers.

It will be shown later in Figs. 3 and 4 that the results of this model are not very accurate in settings we consider. We propose a modification to the model [7] which replaces ρ by

$$\rho_{slot} = \frac{\lambda(\bar{w}_u + \mathbb{E}[R_u])}{S_s(\bar{w}_s + \mathbb{E}[R_s])}, \tag{35}$$

where the numerator is the mean number of slots per second in which an unsaturated source has a packet, and the denominator is the mean total number of system slots per second; S_s and λ are the throughput of a saturated source s and the arrival rate of an unsaturated source u; $\bar{w}_u$ and $\mathbb{E}[R_u]$ are the mean number of backoff slots and attempts that a packet from source u encounters before being successfully sent; and $\bar{w}_s$ and $\mathbb{E}[R_s]$ are the corresponding values for source s. In (35), the service time of source u is not used and hence not 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 networks of N_u identical unsaturated sources sending packets of size l_u with Poisson arrival of rate λ, and N_s identical saturated sources sending packets of size l_s. We varied N_u, N_s, λ and l_u. All sources have the same MAC parameters $\langle CW_{\min} = 32, \eta = 1\rangle$.

a) Scenario 1: The collision probability and throughput of a saturated source, and the collision probability and mean access delay of an unsaturated source are shown in Fig. 3 as functions of N_u, parameterized by N_s. These figures show results from our model as well as from [5], [7] and simulation.

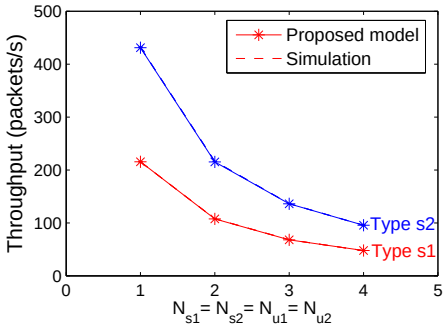
Our model and the model [5] accurately capture the increase in collision probabilities when N_s and N_u increases, and the resulting decrease in throughput and increase in mean access delay. However, collision probabilities and mean access delay from [7] are much higher than those of the simulation.

b) Scenario 2: The collision probability and throughput of each saturated source, and the collision probability and mean access delay of an unsaturated source are shown in Fig. 4 as functions of l_u, parameterized by λ. Results are obtained from our model, [5], [7], and simulation.

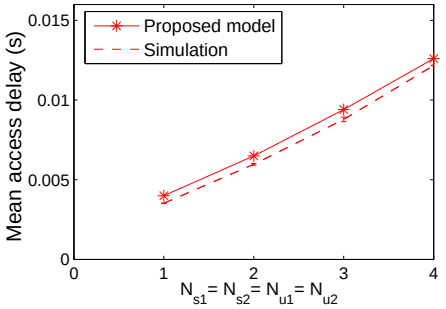
Figure 4 shows that results from our model correctly capture the increase in collision probability with increasing l_u and λ, and the resulting decrease in throughput and increase in mean access delay. As for Scenario 1, the model in [7] overestimates the collision probabilities and mean access delay.

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.



(a) Throughput of saturated sources



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

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

B. Validation in 802.11e EDCA

1) Scenario 3: We simulated networks with 4 traffic types, 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 are distinguished by subscripts $u1$ to $s2$. Unsaturated sources of types $u1$ and $u2$ have arrival rates λ_{u1} and λ_{u2}.

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 $\langle 96, 2\rangle$.

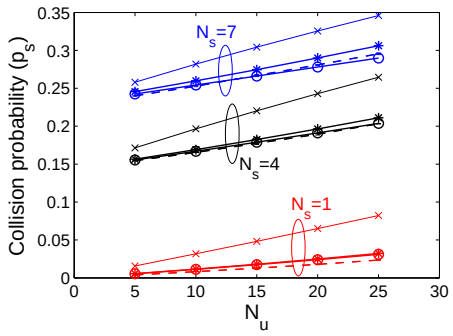
The throughput of a source of type $s1$ and $s2$, and the mean access delay of a source of type $u1$ are shown in Figs. 5(a) and 5(b) as functions of the number of sources per type.

From Fig. 5(a), 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. Our model provides a surprisingly accurate estimate of the throughput.

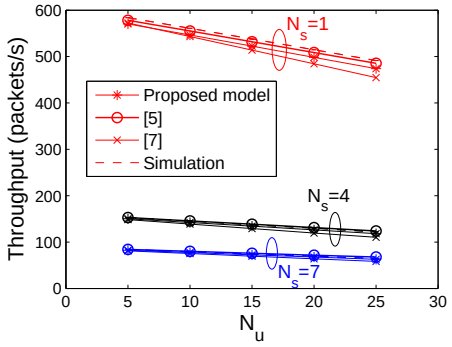
Fig. 5(b) shows that our model provides a reasonably accurate estimate of the mean access delay despite its simplicity compared with Markov chain based models. The model also predicts the access delay of sources of type $u2$ with accuracy similar to that of type $u1$.

2) Scenario 4: We simulated networks of N_u identical unsaturated sources sending bursts of η_u packets of size l_u with the packet arrival rate λ, and N_s identical saturated sources sending fixed bursts of η_s packets of size l_s.

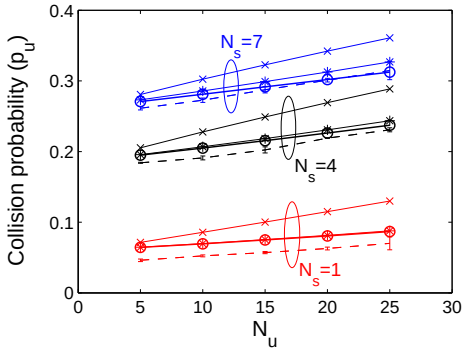
QoS parameters $\langle CW_{\min}, \eta\rangle$ of unsaturated and saturated



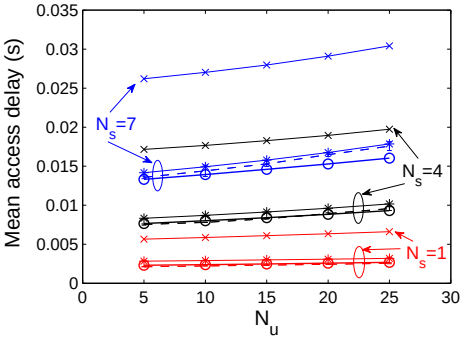
(a) Collision probability of a saturated source



(b) Throughput of a saturated source

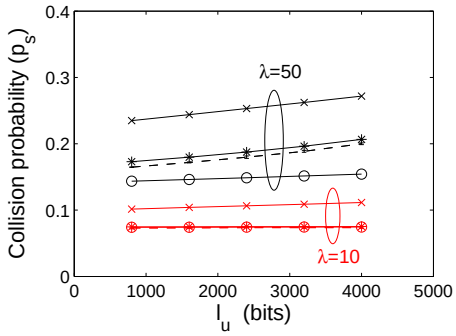


(c) Collision probability of an unsaturated source

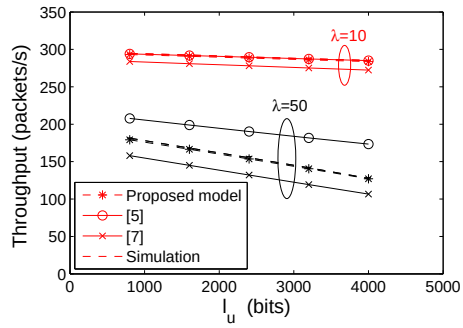


(d) Mean access delay of an unsaturated source

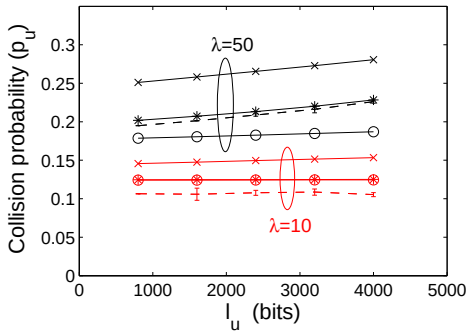
Fig. 3. Collision probabilities, throughput, and mean access delay for DCF, Scenario 1. Figs. 3(a), 3(c) and 3(d) clearly show that our model is much more accurate than the model in [7]. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u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



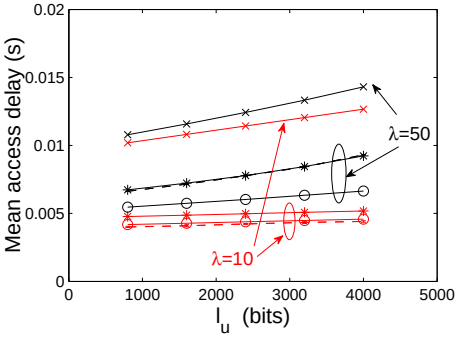
(a) Collision probability of a saturated source



(b) Throughput of a saturated source

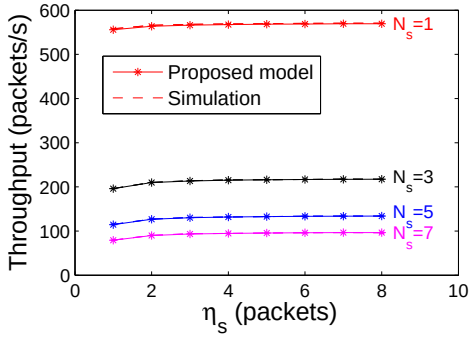


(c) Collision probability of an unsaturated source

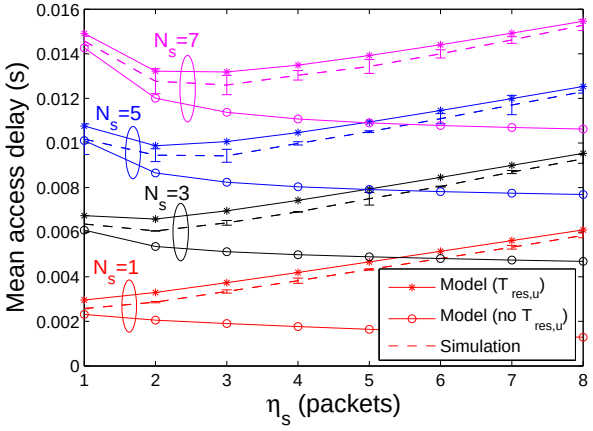


(d) Mean access delay of an unsaturated source

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, $W_s = 32$, $\eta_s = 1$.)



(a) Throughput of a saturated source.



(b) Mean access delay of an unsaturated source

Fig. 6. 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, $W_s = \eta_s W_u$.)

sources, respectively, are $\langle 32, 1\rangle$ and $\langle 32\eta_s, \eta_s\rangle$.

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 [32]), 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 in packets/s of a saturated source is shown in Fig. 6(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 increases.

One of our model's contributions is to capture the residual time of busy period during which a burst arrived $T_{\text{res},u}$, which was not important in DCF and has often been overlooked in EDCA models. Fig. 6(b) shows the mean access delay of a burst from unsaturated sources with and without $T_{\text{res},u}$ in the access delay models under the same scenario. As seen, when η_s is large, $T_{\text{res},u}$ has significant effect on delay estimation.

Also from Fig. 6(b), when η_s increases, for $N_s > 1$, 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 saturated sources. For larger η_s, the increase in residual time $T_{\text{res},u}$ dominates this. This suggests 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. 6 shows that increasing W_s and η_s together can benefit both unsaturated 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. Our model 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, here we approximate K and m to be infinite in the whole model and $b_u = 1$ 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

Note that access delay distribution can be calculated using transform methods. The generating function of 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 the Lattice-Poisson algorithm [27]. The details are not illuminating and hence referred to [23].

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

$$W_{\text{med}}(k) = \sum_{j=0}^{k} M[U_{uj}] = \left(2^k - \frac{1}{2}\right) W_u - \frac{k+1}{2}. \quad (36)$$

Note 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

$$\begin{aligned} ccdf_W(W_{\text{med}}(k)) &\le 1 - \left(\sum_{j=0}^{k-1}(1-p_u)p_u^j + \frac{1}{2}(1-p_u)p_u^k\right) \\ &= p_u^k\left(\frac{1+p_u}{2}\right), \end{aligned} \quad (37)$$

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

$$\begin{aligned} D(W_{\text{med}}(k)) &\approx W_{\text{med}}(k)\mathbb{E}[Y_u] + k\mathbb{E}[C_u] + \mathbb{E}[T_{\text{res},u}] + \mathbb{E}[T_u^s] \\ &= 2^k W_u \mathbb{E}[Y_u] + k(\mathbb{E}[C_u] - \mathbb{E}[Y_u]/2) + K \equiv f(k). \end{aligned} \tag{38}$$

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,

$$ccdf_D(D) \approx \left(\frac{1+p_u}{2}\right) p_u^{f^{-1}(D)}. \tag{39}$$

It turns out that (39) is a good approximation for any delay $D \geq D(W_{\text{med}}(0))$.

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

$$\begin{aligned} ccdf_D(D(j)) &\approx 1 - (1-p_u)\frac{j+1}{W_u} \\ &= 1 - \frac{1-p_u}{W_u}\Big(1 + \frac{D(j) - \mathbb{E}[T_{\text{res},u}] - \mathbb{E}[T_u^s]}{\mathbb{E}[Y_u]}\Big). \end{aligned} \tag{40}$$

Thus, we propose the approximation that finds the ccdf from (40) for delays less than $D((W_u - 1)/2)$, and from (39) 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 ($At^k P(D > t) \to 1$ as $t \to \infty$ for some A, k). Although the true delay cannot be strictly heavy tailed when retry limit is finite, the approximation holds for delays in the typical range of interest, from 10 ms to 1 s [33].

This power law arises since the duration and probability of occurrence of the kth backoff stage increase geometrically in k. This is distinct from the heavy tailed delays in ALOHA, 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, 802.11 can be usefully modeled as heavy tailed even with typical limits of 6 to 8 retransmissions.

Note from (38) 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 (39), the complementary CDF of a large delay D is approximately

$$ccdf_D(D) \approx \frac{1+p_u}{2}\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)^{\log_2(p_u)}. \tag{41}$$

That is, the distribution has 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 [34]. 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 BEB. If each source is assumed to implement an M/G/1 queue, then this corresponds to the service time having infinite variance.

Consider a log-log plot of the ccdf of a random variable D whose ccdf is the right hand side of (41). The minimum (steepest) slope for which the variance of D becomes infinite is -2 [34]. The right hand side of (41) 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 (11) and (33)–(34), 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 using the same backoff parameters as saturated sources will have $p_u \geq 1/4$. Let us start with the following lemma, proved in Appendix A.

Lemma 1: Let s and u denote an arbitrary saturated and unsaturated source. Under the model (11) and (33),

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

If, in addition, (34) holds then $p_u > p_s$.

Theorem 2: Consider the model (11) and (33)–(34), with all sources using the same backoff parameters ($W_x = W, \forall x \in \mathbb{S} \cup \mathbb{U}$). If

$$N_s \geq 1 + \frac{\log(3/4)}{\log(1 - \frac{4}{3W+2})} \tag{42}$$

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 (41) is infinite. The proof is in Appendix A. Surprisingly, the sufficient condition for infeasibility (42) depends only on W, the minimum contention window, and not settings such as channel data rate, traffic of real-time source, or the *TXOP limit*.

From (41), the distribution of an unsaturated source's access delay D_u under the model (11)–(34) has a tail which is approximately power law, given by the right hand side of (41). Hence, under the condition (42), the variance of the 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 is to validate: (i) approximation method of determining access delay distribution; (ii) the slope of the distribution curve's tail; (iii) the condition (42) for the infinite variance of unsaturated sources' access delay.

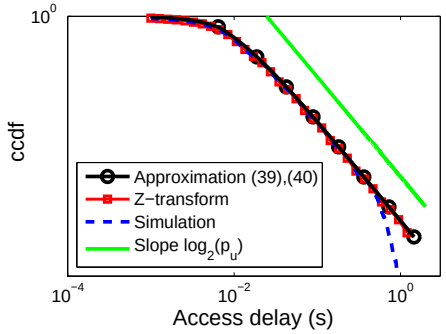


Fig. 7. Distribution of access delay. (Unsaturated stations: Poisson arrivals with rate $\lambda = 10$ packets/s, $N_u = 20$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 6$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

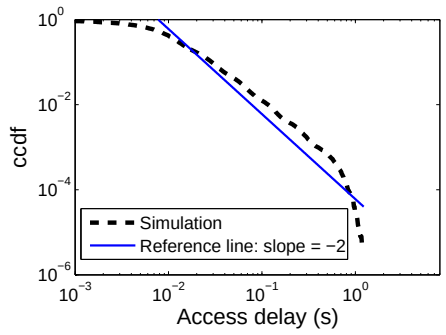


Fig. 8. Access delay distribution of unsaturated sources. (Unsaturated stations: Poisson with $\lambda = 10$ packets/s, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 8$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

The simulated network is the same as that in Section IV. In the simulation, 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 and simulation are shown in Fig. 7. Although 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. The approximation is of comparable accuracy to the z-transform method.

2) Slope of distribution curve's tail: The straight line in Fig. 7 shows the slope $\log_2(p_u)$. It captures the trend of the distribution curve reasonably well in the typical delay range from tens to hundreds of ms.

3) Validation of Theorem 2: From (42), when W is 32 as in 802.11 DCF, the minimum number of saturated sources required for infinite variance of unsaturated sources' access delay is 8. This is validated in Fig. 8 which shows the access delay distribution of unsaturated sources from NS-2 simulation. As 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 implies that these delays will occur as often as if the system had a power law tail with infinite variance.

VI. Conclusion

We have provided a comprehensive but tractable fixed point model of 802.11 WLANs with 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 unsaturated sources, which is sufficiently accurate at low queue occupancies to predict the burst size distribution.

Using the model to investigate the interaction between these two traffic types, 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. From this, the slope $\log_2(p_u)$ of distribution curve's tail has been obtained and 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
Proof of Theorem 2

Proof of Lemma 1: Dividing p_s from (11c) by p_u from (11c), we have

$$\frac{1-p_u}{1-p_s} = \frac{1-\tau_s}{1-\tau_u}. \tag{43}$$

Moreover, by (33),

$$\tau_s = \frac{S_s}{\mathbb{E}[\eta_s]} \frac{\mathbb{E}[Y]}{1-p_s}. \tag{44}$$

Dividing (44) by τ_u from (11b), and applying (43) gives

$$\frac{\tau_s}{\tau_u} = \frac{S_s \mathbb{E}[\eta_u]}{\lambda_u \mathbb{E}[\eta_s]} \frac{1-p_u}{1-p_s} = \frac{S_s \mathbb{E}[\eta_u]}{\lambda_u \mathbb{E}[\eta_s]} \frac{1-\tau_s}{1-\tau_u} \tag{45}$$

which establishes the first claim.

By (34), this implies $\tau_s > \tau_u$, whence $p_u > p_s$ by (43). ■

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
$$N_s \geq 1 + \frac{\log(3/4)}{\log(1 - \frac{4}{3W+2})} \tag{46}$$
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 (41) is infinite.

These can be shown as follows:

1) This follows from (11c) since $\tau_u \in [0,1]$, and τ_s is decreasing in p_s.
2) When $N_u = 0$, (11c) becomes $p_s = 1 - (1-\tau_s)^{N_s - 1}$. Thus $p_s \geq 1/4$ if
$$\tau_s \geq 1 - \left(\frac{3}{4}\right)^{1/(N_s-1)}. \tag{47}$$
Conversely, (11a) decreases in p_s, and so $p_s \geq 1/4$ if
$$\tau_s \leq \frac{4}{3W+2}. \tag{48}$$

Combining (47) and (48), $p_s \geq 1/4$ if

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

which upon rearrangement gives (46).

3) If $p_u > 1/4$, then the random variable whose ccdf is the right hand side of (41) has a tail heavier than kD_u^{-2} for some k, and hence infinite variance.

■

ACKNOWLEDGMENT

This work was supported by Australian Research Council grants DP1095103 and FT0991594.

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*, 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 CCNC*, 2007.

[8] I. Inan, F. Keceli, and E. Ayanoglu, "Modeling the 802.11e Enhanced Distributed Channel Access Function," *Proc. IEEE GLOBECOM*, 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 ICIEA*, 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 ICC*, 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*, 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.*, 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, 2005.

[16] B. Bellalta, C. Cano, M. Oliver, and M. Meo, "Modeling the IEEE 802.11e EDCA for MAC parameter optimization," *Proc. IEEE CCNC*, 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 ICMMC*, 2007.

[18] S. H. Nguyen, H. L. Vu and L. L. H. Andrew, "Service differentiation without prioritization in 802.11 WLANs," *Proc. IEEE LCN*, 2011.

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

[20] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew, "Packet size variability affects collisions and energy efficiency in WLANs," *Proc. IEEE WCNC*, 2010.

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

[22] 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, 2009.

[23] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew, "Performance analysis of 802.11 WLANs with saturated and unsaturated sources," Available ⟨http://caia.swin.edu.au/reports/110811A/CAIA-TR-110811A.pdf⟩.

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

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

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

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

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

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

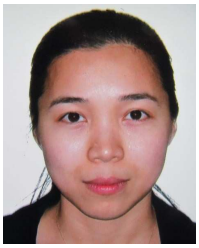
[30] "The network simulator ns-2," Available at http://www.isi.edu/nsnam/ns/.

[31] 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/.

[32] 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.

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

[34] 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.



Suong H. Nguyen received B.Sc. degree and M.Sc. from the Post and Telecommunications Institute of Technology, Vietnam in 2000 and La Trobe University, Australia in 2007, respectively. She is currently Ph.D. student in Swinburne University of Technology, Australia. Her research interest includes wireless communication and optical transmission system.

Hai L. Vu (S'97M'98-SM'06) received the B.Sc./M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Budapest, Budapest, Hungary, in 1994 and 1999, respectively. From 1994 to 2000, he was a Research Engineer with Siemens AG, Hungary. During this period, his focus was on performance measurements, Internet quality of service, and IP over ATM. During 2000-2005, he was with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia. In 2005, he joined Swinburne University of Technology and is with the Centre for Advanced Internet Architectures (CAIA). He is currently an Associate Professor at the Faculty of Information and Communication Technologies (FICT), Swinburne University of Technology, Hawthorn, Victoria, Australia. Dr. Vu has authored or coauthored over 100 scientific journals and conference papers. His research interests include performance analysis and design of wireless data networks, and stochastic optimization with applications to Intelligent Transport Systems (ITS) and SmartGrid.

Lachlan Andrew (M'97-SM'05) received the B.Sc., B.E. and Ph.D. degrees in 1992, 1993, and 1997, from the University of Melbourne, Australia. Since 2008, he has been an associate professor at Swinburne University of Technology, Australia, and since 2010 he has been an ARC Future Fellow. From 2005 to 2008, he was a senior research engineer in the Department of Computer Science at Caltech. Prior to that, he was a senior research fellow at the University of Melbourne and a lecturer at RMIT, Australia. His research interests include energy-efficient networking and performance analysis of resource allocation algorithms. He was co-recipient of the best paper award at IEEE INFOCOM 2011 and IEEE MASS 2007. He is a member of the ACM.