MAC Access Delay of IEEE 802.11 DCF

Taka Sakurai, *Member, IEEE,* and Hai L. Vu, *Senior Member, IEEE*

***Abstract*— The MAC access delay in a saturated IEEE 802.11 DCF wireless LAN is analyzed. We develop a unified analytical model and obtain explicit expressions for the first two moments as well as the generating function. We show via comparison with simulation that our model accurately predicts the mean, standard deviation, and distribution of the access delay for a wide range of operating conditions. In addition, we show that the obtained generating function is much more accurate than others that have appeared in the literature.**

Using our model, we prove that the binary exponential backoff mechanism induces a heavy-tailed delay distribution for the case of unlimited retransmissions. We show using numerical examples that the distribution has a truncated power-law tail when a retransmission limit exists. This finding suggests that DCF is prone to long delays and not suited to carrying delay-sensitive applications.

***Index Terms*— Medium access delay, IEEE 802.11, wireless LAN, performance analysis, generating function, heavy tail.**

I. Introduction

In recent years, demand for wireless Internet connectivity has led to a proliferation of wireless local area networks (WLANs). Products based on the IEEE 802.11 family of standards have acquired the lion's share of this burgeoning market. As in the wider Internet, the majority of traffic carried on a typical IEEE 802.11 WLAN today consists of non-realtime applications such as web browsing and email. However, in the near future, it is expected that a significant proportion of the traffic on WLANs will consist of realtime applications such as voice over IP. To understand the potential for IEEE 802.11 WLANs to support such delay-sensitive applications, performance models for evaluating delay characteristics are needed. This paper is a step in this direction. We study the *access delay* of the MAC layer when there are multiple wireless stations in ideal channel conditions, where each station always has a packet available for transmission.

Formally, we define the access delay as the time interval between the instant when the packet reaches the head of the transmission queue and begins contending for the channel, and the time when the packet is successfully received at the destination station. The IEEE 802.11 MAC layer employs a channel access mechanism called the distributed coordination function (DCF). DCF is a backoff protocol, so the access delay is a stochastic quantity.

The analysis of a buffered backoff protocol system amounts to the analysis of a system of coupled queues [1]. Exact delay analysis of such systems has so far proved elusive, so approximation techniques are typically used. The approximate analysis of the access delay (or the closely related inter-departure time) in IEEE 802.11 WLANs has been the subject of several papers. While most studies consider only the mean delay, Carvalho and Garcia-Luna-Aceves [2] find approximate formulae for both the mean and variance of the inter-departure time by working with distributions. However, their model assumes features inconsistent with DCF, such as infinite retransmissions; moreover, simulation results presented in the same paper reveal that the formulae lack accuracy. In a recent paper, Zhai, Kwon and Fang [3] derive the generating function of the probability mass function of the inter-departure time. The generating function is derived from an approximate Markov chain model of the DCF introduced in the seminal paper by Bianchi [4]. Tickoo and Sikdar [5] derive a different expression for the generating function of the inter-departure time using probabilistic arguments.

In this paper, we build a detailed stochastic model of DCF to obtain an approximate yet accurate expression for the access delay random variable. The expression enables a unified analysis of the moments and the generating function. We derive explicit formulae for the mean, standard deviation, and generating function. In principle, it should be possible to obtain all moments of the access delay by repeated differentiation of the generating function followed by appropriate limit taking, which is the approach suggested in [5] and [3]. However, the generating function in question is complicated, making this approach extremely tedious, so the explicit moment expressions we obtain have utility. We demonstrate that numerical transform inversion [6] can be used to obtain values of the distribution from the generating function. With the aid of the ns-2 simulator [7], we show that our analytical formulae for the mean and standard deviation are very accurate. Significantly, we also find that the distribution values obtained by numerical inversion are in excellent agreement with the simulation results. Our generating function differs from that in [5] and [3], and we show that numerical inversion of these other generating functions leads to inaccurate distributional values.

Another major contribution of this paper is an asymptotic analysis of the model for the theoretical setting of unlimited retransmissions. We prove that the distribution of the access delay is heavy-tailed and that, as the number of active stations increases, the number of bounded moments decreases. The analysis shows that the asymptote of the mean access delay

Manuscript received June 23, 2005; revised September 13, 2006; accepted January 29, 2007. The associate editor coordinating the review of this letter and approving it for publication was K. K. Leung. This work was presented in part at the 19th International Teletraffic Congress (ITC), Beijing, 2005.

T. Sakurai is with the ARC Special Research Centre for Ultra-Broadband Information Networks (CUBIN), Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia. CUBIN is an affiliated program of National ICT Australia (email: tsakurai@ieee.org).

H. L. Vu is with the Centre for Advanced Internet Architectures (CAIA), Faculty of I.C.T., Swinburne University of Technology, Hawthorn, VIC 3122, Australia (email: h.vu@ieee.org).

Digital Object Identifier 10.1109/TWC.2007.05458.

1536-1276/07$25.00 © 2007 IEEE

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

is linear in the number of active stations. Through numerical examples, we show that the heavy-tailed characteristics in the theoretical system manifest as truncated heavy-tailed characteristics in systems with limited retransmissions i.e. DCF. The access delay distributions exhibit truncated power-law tails, the mean delay curves are initially close to linear, and the standard deviation curves grow rapidly with both the number of stations and the retransmission limit. These findings suggest that DCF is poorly suited to carrying delay-sensitive applications.

The rest of this paper is organized as follows. In Section II, we briefly describe the IEEE 802.11 Medium Access Control (MAC) protocol. In Section III, we present our analysis of the access delay. We first review a fixed point approximation for the packet collision probability which was introduced by Bianchi [4]. Then, we develop a probabilistic expression for the access delay that we exploit to obtain the mean, standard deviation and generating function. In Section IV, we present an asymptotic analysis of our model and in Section V, we compare our analytical results with simulation and explore the implications of our asymptotic analysis. Finally, Section VI contains concluding remarks.

II. The 802.11 MAC protocol

The IEEE 802.11 standard [8] specifies a multiple access mechanism called the distributed coordination function (DCF), in which nodes contend for the channel using a carrier sense multiple access mechanism with collision avoidance (CSMA/CA). To reduce the incidence of collisions, DCF employs both sensing of the channel to detect channel activity and truncated binary exponential backoff (BEB) to randomize the start times of packet transmissions.

After every successful data packet transmission, a station initiates a post-transmission random backoff. If the next packet was already enqueued when the previous packet was sent, its defer time will span the entire backoff period, whereas a packet that arrives at the MAC layer after the previous packet was sent would experience only part of the backoff period, or none at all if the backoff period has already elapsed. We limit attention to the scenario when stations always have packets backlogged — the so-called 'saturated' case — so every packet defers for the entire post-transmission backoff period associated with the previous packet.

Prior to a backoff interval, the channel must be sensed idle for a guard period known as the distributed interframe space, DIFS. Backoff intervals are slotted, and stations are only permitted to commence transmissions at the beginning of slots. When backoff is initiated, a random backoff time is selected, representing the number of idle slots that must pass before the next packet can be transmitted. The discrete backoff time is uniformly distributed in the range $[0, CW-1]$, where CW is the contention window. At the first transmission attempt, CW is set equal to W, the minimum contention window. The backoff time counter is decremented by one at the end of each idle slot. It is *frozen* when a packet transmission is detected on the channel, and reactivated after the channel is sensed idle again for a guard period. The guard period is equal to a DIFS if the transmitted packet was error-free, and equal to the extended interframe space, EIFS, if the packet was in error. The station transmits when the backoff time counter reaches zero. A collision occurs when the counters of two or more stations reach zero in the same slot.

We digress briefly to touch upon a little-known subtlety of DCF behaviour. The backoff operation described above induces a significantly lower level of contention in the slot immediately following a DIFS compared to other slots [9], since the only stations that attempt in this slot are those that transmitted in the preceding busy period *and* selected a new backoff time of zero. This phenomenon, which we dub the *reduced contention effect*, presents a complication for accurate analysis of the collision probability, and will be discussed further in Sections III and V.

Following a successful packet transmission, the receiving MAC layer sends an ACK after a short interframe space, SIFS, where a SIFS is shorter than a DIFS. This ordering relationship ensures that ACK packets are sent without contention. If the packet transmission is unsuccessful (an event indicated by an ACK timeout at the sending station), the congestion window evolves according to truncated BEB. The congestion window size is multiplied by 2, and another backoff period is initiated. Window doubling continues until the maximum possible value, $CW_{max} = 2^m W$, is reached. If the packet is unsuccessful after m attempts, the window is maintained at CW_{max} for the remaining attempts until the packet is successful, or until the maximum number of attempts, K, is reached. If the packet is still unsuccessful after K attempts, the packet is discarded and the MAC protocol reports back to the layer above. In this paper, we are only interested in the delay of packets that are successful at the MAC layer.

The two-way handshaking scheme described above, where the sender transmits a MAC data packet and the receiver responds with a MAC ACK, is known as the basic access mechanism of DCF. Another access mechanism defined for DCF is a four-way handshaking scheme called RTS/CTS. While we restrict our attention to the basic access mechanism setting, our model can be readily extended to the RTS/CTS mechanism.

III. Analysis of Access Delay

In our study, we consider a population of n stations, each operating in saturation, and assume ideal channel conditions (no channel errors, hidden terminals or capture). While DCF employs truncated BEB, we develop our analysis for the more general framework of truncated exponential backoff (EB) with backoff multiplier $\lambda \geq 1$.

A. Collision Probability

An important ingredient for our delay analysis is the collision probability p seen by a packet transmitted on the channel. A fixed point formulation for p was introduced by Bianchi [4], who proposed the relationship

$$p = 1 - (1-\tau)^{n-1}, \tag{1}$$

where τ is the attempt probability and is a function of p. Bianchi used a two-dimensional Markov chain analysis to obtain an expression for τ for the case of no retransmission

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

limit. Later, the retransmission limit was addressed in [10]. Instead of a two-dimensional Markov chain analysis, Kwak, Song and Miller [11] used a simpler analysis based on a one-dimensional Markov chain. Finally, Kumar et al. [12] recently developed an expression for τ using a renewal theory approach.

In our work, we adopt the approach proposed in [11], for which we provide a brief summary here. The idea is to obtain a mean-value analysis of the evolution of a node's backoff process over the K backoff periods. According to the protocol description in Section II, the backoff period durations $U^{(i)}$, $i = 0, \ldots, K-1$, are discrete uniform random variables given by

$$U^{(i)} = \begin{cases} \mathcal{U}(0, \lambda^i W - 1), & i = 0, \ldots, m-1, \\ \mathcal{U}(0, \lambda^m W - 1), & i = m, \ldots, K-1, \end{cases} \tag{2}$$

where $\mathcal{U}$ is the uniform distribution. From (2), it can be shown that the average backoff durations are

$$\mathrm{E}[U^{(i)}] = \begin{cases} (\lambda^i W - 1)/2 & \text{for } i = 0, \ldots, m-1, \\ (\lambda^m W - 1)/2 & \text{for } i = m, \ldots, K-1. \end{cases} \tag{3}$$

Let π_i denote the relative frequency that a node enters the ith backoff period in steady state. In [11], it is shown that $\pi_i = (1-p)p^i(1-p^K)^{-1}$ and that the reciprocal of the attempt probability is

$$\begin{aligned} \tau^{-1} &= \sum_{i=0}^{K} \pi_i \, \mathrm{E}[U^{(i)}] \\ &= \frac{(1-p)W(1-(\lambda p)^m)}{2(1-p^K)(1-\lambda p)} + \frac{\lambda^m W(p^m - p^K)}{2(1-p^K)} \\ &\quad - \frac{1}{2}. \end{aligned} \tag{4}$$

Equation (4) has the intuitive interpretation that the reciprocal of the attempt probability is equal to the mean backoff time. It can be shown that the approach of [12] also leads to (4). Equations (1) and (4) establish a fixed point formulation from which p can be computed using a numerical technique.

The starting assumption of our analysis, and all analyses for p cited above, is that p is the same for every slot, which clearly ignores the reduced contention effect. We investigate this effect in Section V through a numerical example. We will find, however, that this does not have a significant impact on the accuracy of our delay analysis.

B. Expression for the Access Delay

We consider a selected (tagged) station and derive an expression for the access delay as seen by packets of this station under saturation. From the protocol description in Section II, we can identify several events that contribute to the access delay. The most obvious is the successful transmission of the packet. Preceding this event will be the initial (post-transmission) backoff plus a variable number of collisions involving the tagged station and subsequent backoff periods. Successful transmissions and collisions not involving the tagged station also contribute to the access delay, since they manifest as interrupts to the backoff counter. The access delay of a packet can be large if there are many collisions and interrupts to the backoff timer. However, it is important to recognize that the access delay is always bounded since the number of retransmissions is limited.

Let D be the random variable (r.v.) representing the access delay. We write

$$D = A + T, \tag{5}$$

where T is a r.v. representing the channel occupancy of a transmitted packet. The r.v. A represents the sum of durations of collisions and backoffs involving the tagged station, and the durations of successful transmissions and collisions by non-tagged stations. Since the number of backoff periods depends on the number of retransmissions, the value of A strongly depends on the number of retransmissions. We therefore deduce that the distribution of A can be represented as a simple mixture of distributions, where each component distribution represents a conditional distribution *conditioned* on a certain number of retransmissions taking place, and the mixing probabilities are the respective probabilities of the number of retransmissions. The number of retransmissions before success obeys a truncated geometric distribution, so that the probability of i retransmissions is ηp^i, where $\eta = (1-p)(1-p^K)^{-1}$. We therefore have that

$$A = A^{(i)} \quad \text{w.p.} \quad \eta p^i, \quad 0 \le i \le K-1, \tag{6}$$

where 'w.p.' means 'with probability'. The generic component r.v. $A^{(i)}$ is comprised of i collisions, $i+1$ backoff intervals, and their associated interruptions. We write

$$A^{(i)} = \sum_{j=0}^{i} B_i^{(j)} + \sum_{j=1}^{i} C_{ij}, \tag{7}$$

where it is understood that if $i = 0$, the value of the second sum is zero. The r.v.'s C_{ij} account for the channel occupancies of collisions involving the tagged user, while the $B_i^{(j)}$ represent the backoff intervals and their interruptions. Clearly, the r.v.'s C_{ij}, $i = 1, \ldots, K-1$ and $j = 1, \ldots, i$, are i.i.d.. The r.v.'s $B_i^{(j)}$, $i = 0, \ldots, K-1$ and $j = 0, \ldots, i$, are independent and, for fixed j, are i.i.d. in i.

The scope of $B^{(j)}$ (for simplicity, we drop the index i from the notation) is defined by a backoff interval that takes a discrete uniform distribution. Each slot of the backoff interval can, with certain probabilities, be interrupted at the start of the slot by a successful transmission by a station other than the tagged one, or by a collision not involving the tagged station. After an interruption, the backoff slot that was interrupted is assumed to pass without further incident. In other words, we ignore the possibility of the same slot being interrupted more than once. Our justification for making this simplification is that, due to the reduced contention effect, the probability of multiple interruptions is small.

The preceding arguments imply that $B^{(j)}$ can be represented as a random sum:

$$B^{(j)} = \sum_{k=1}^{U^{(j)}} (t_{slot} + Y_k), \tag{8}$$

where t_{slot} is the duration of a slot, Y_k are i.i.d. and represent interruptions, and $U^{(j)}$ are defined in (2).

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

It remains to decompose Y_k into its constituent parts. In the following, we denote a generic Y_k by Y. The random variable Y is equal to zero if none of the non-tagged stations transmit in the slot. If there is only one transmission, Y is equal to the channel occupancy of a successful transmission as seen from the point of view of the tagged station, and if there is more than one transmission, Y equals the channel occupancy of a collision as seen from the point of view of the tagged station. It can be shown that the probability of at least one transmission by the $n-1$ non-tagged stations is the same as the collision probability p, and the probability q of only one transmission is

$$q = (n-1)\tau(1-\tau)^{n-2}. \tag{9}$$

We can now write Y as

$$Y = \begin{cases} 0 & \text{w.p.} \quad 1-p, \\ T^* & \text{w.p.} \quad q, \\ C^* & \text{w.p.} \quad p-q, \end{cases} \tag{10}$$

where T^* is a r.v. representing the channel occupancy of a successful transmission of a station other than the tagged station, and C^* is a r.v. representing the channel occupancy of a collision not involving the tagged station.

In our analysis, we permit general distributions for the occupancy r.v.'s T, T^*, C and C^*, which is tantamount to allowing general distributions for the data packet lengths. However, for the numerical examples in Section V, we assume constant length data packets for all stations for simplicity. Let t_{data} denote the fixed transmission time of a data packet, and t_{ack} be the transmission time of the ACK packet. Also denote the duration of the $DIFS$ and $SIFS$ by t_{difs} and t_{sifs}, respectively. We have

$$T = t_{data} + t_{difs}, \tag{11}$$

$$C = T^* = C^* = t_{data} + t_{sifs} + t_{ack} + t_{difs}, \tag{12}$$

where C denotes a generic C_{ij}.

C. Mean and standard deviation

We now derive the mean $\mathrm{E}[D]$ and standard deviation $\mathrm{StdDev}[D]$ of the access delay from the analysis of the previous section. From (5), we trivially obtain

$$\mathrm{E}[D] = \mathrm{E}[A] + \mathrm{E}[T], \tag{13}$$

$$\mathrm{StdDev}[D] = (\mathrm{Var}[A] + \mathrm{Var}[T])^{1/2}, \tag{14}$$

where $\mathrm{Var}[.]$ denotes the variance.

From (6), we obtain

$$\mathrm{E}[A] = \eta \sum_{i=0}^{K-1} p^i \,\mathrm{E}[A^{(i)}],$$

$$\mathrm{Var}[A] = \eta \sum_{i=0}^{K-1} p^i (\mathrm{Var}[A^{(i)}] + (\mathrm{E}[A^{(i)}] - \mathrm{E}[A])^2).$$

The mean and variance of $A^{(i)}$, are derived from (7):

$$\mathrm{E}[A^{(i)}] = \sum_{j=0}^{i} \mathrm{E}[B^{(j)}] + i\,\mathrm{E}[C],$$

$$\mathrm{Var}[A^{(i)}] = \sum_{j=0}^{i} \mathrm{Var}[B^{(j)}] + i\,\mathrm{Var}[C].$$

Next, we derive the mean and variance of $B^{(j)}$. From (8), it follows that

$$\mathrm{E}[B^{(j)}] = \theta\,\mathrm{E}[U^{(j)}],$$

$$\mathrm{Var}[B^{(j)}] = \mathrm{E}[U^{(j)}]\,\mathrm{Var}[Y] + \theta^2\,\mathrm{Var}[U^{(j)}],$$

where $\theta = t_{slot} + \mathrm{E}[Y]$. The mean $\mathrm{E}[U^{(j)}]$ of the uniform distribution is given by (3), and it can be shown from (2) that

$$\mathrm{Var}[U^{(j)}] = \begin{cases} (\lambda^{2j}W^2 - 1)/12 & \text{for } j = 0, \ldots, m-1, \\ (\lambda^{2m}W^2 - 1)/12 & \text{for } j = m, \ldots, K-1. \end{cases} \tag{15}$$

Next, from (10), we obtain

$$\mathrm{E}[Y] = q\,\mathrm{E}[T^*] + (p-q)\,\mathrm{E}[C^*], \tag{16}$$

$$\begin{aligned}\mathrm{Var}[Y] = {} & q(\mathrm{Var}[T^*] + (\mathrm{E}[T^*] - \mathrm{E}[Y])^2) \\ & + (p-q)(\mathrm{Var}[C^*] + (\mathrm{E}[C^*] - \mathrm{E}[Y])^2).\end{aligned} \tag{17}$$

Putting everything together, we finally obtain

$$E[D] = \eta \sum_{i=0}^{K-1} p^i \left\{ (t_{slot} + E[Y]) \sum_{j=0}^{i} E[U^{(j)}] + iE[C] \right\} + E[T], \tag{18}$$

$$\begin{aligned}\mathrm{Var}[D] = {} & \eta \sum_{i=0}^{K-1} p^i \left\{ \sum_{j=0}^{i} \left(E[U^{(j)}]\,\mathrm{Var}[Y] + \theta^2\,\mathrm{Var}[U^{(j)}] \right) \right. \\ & \left. + i\,\mathrm{Var}[C] + \left(\theta \sum_{j=0}^{i} E[U^{(j)}] + iE[C] - E[A] \right)^2 \right\} \\ & + \mathrm{Var}[T],\end{aligned} \tag{19}$$

where $E[U^{(j)}]$,$\mathrm{Var}[U^{(j)}]$, $E[Y]$ and $\mathrm{Var}[Y]$ are given by (3), (15), (16) and (17), respectively.

For the case of constant length data packets, we obtain from (11) and (12),

$$\begin{aligned} E[T] &= t_{data} + t_{difs}, \\ E[C] &= E[T^*] = E[C^*] = t_{data} + t_{sifs} + t_{ack} + t_{difs}, \\ Var[T] &= Var[C] = Var[T^*] = Var[C^*] = 0. \end{aligned}$$

D. Generating function

We adopt the following notational convention for a generating function: if X is a non-negative, integer-valued random variable, then we denote the generating function of the probability mass function (pmf) of X by

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

All the random variables described in Section III-B were non-negative and discrete, but not necessarily integer-valued. However, they can be easily transformed to integer-valued random variables by defining a lattice, with spacing δ say, such that the values of all random variables fall on the lattice points, and then scaling δ to one. In the following analysis, we avoid an additional set of notation for the scaled random

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

variables by reusing the random variable names in Section III-B to refer to their integer-valued correspondents.

We commence our analysis by writing down an expression for $\widehat{D}(z)$ based on (5):

$$\widehat{D}(z) = \widehat{A}(z)\widehat{T}(z).$$

In the numerical examples presented in Section V, we focus on the complementary cumulative distribution function (CCDF) of the access delay rather than the pmf. The generating function of the CCDF, $\widehat{D_c}(z)$, can be obtained from $\widehat{D}(z)$ via the identity

$$\widehat{D_c}(z) = \frac{1-\widehat{D}(z)}{1-z}. \tag{20}$$

The next task is to find an expression for $\widehat{A}(z)$. From (6), it can be shown that

$$\widehat{A}(z) = \eta \sum_{i=0}^{K-1} p^i \widehat{A^{(i)}}(z),$$

and from (7), we obtain

$$\widehat{A^{(i)}}(z) = \widehat{C}(z)^i \prod_{j=0}^{i} \widehat{B^{(j)}}(z).$$

From (8), it follows that

$$\widehat{B^{(j)}}(z) = \widehat{U^{(j)}}(z^\sigma \widehat{Y}(z)),$$

where σ is an integer constant defined by $\sigma = t_{slot}/\delta$, with δ being the lattice spacing. Equation (2) yields

$$\widehat{U^{(j)}}(z) = \begin{cases} \frac{1-z^{\kappa(j)}}{\kappa(j)(1-z)} & \text{for } j = 0, \ldots, m-1, \\ \frac{1-z^{\kappa(m)}}{\kappa(m)(1-z)} & \text{for } j = m, \ldots, K-1, \end{cases} \tag{21}$$

where $\kappa(j) = \lambda^j W$.

The remaining task is to derive an expression for $\widehat{Y}(z)$. From (10), it follows that

$$\widehat{Y}(z) = q\widehat{T^*}(z) + (p-q)\widehat{C^*}(z) + (1-p). \tag{22}$$

Thus, the generating function of the pmf of the access delay is given by

$$\widehat{D}(z) = \eta\, \widehat{T}(z) \sum_{i=0}^{K-1} p^i \widehat{C}(z)^i \prod_{j=0}^{i} \widehat{U^{(j)}}(z^\sigma \widehat{Y}(z)), \tag{23}$$

where $\widehat{U^{(j)}}(z)$ and $\widehat{Y}(z)$ are given by (21) and (22), respectively.

For the case of constant length data packets, it follows from (11) and (12) that

$$\widehat{T}(z) = z^\alpha,$$
$$\widehat{C}(z) = \widehat{T^*}(z) = \widehat{C^*}(z) = z^\beta,$$

where α and β are integer constants defined by

$$\alpha = (t_{difs} + t_{data})/\delta,$$
$$\beta = (t_{data} + t_{sifs} + t_{ack} + t_{difs})/\delta.$$

Our analysis approach and obtained generating function differ from that of Tickoo and Sikdar [5] and Zhai, Kwon and Fang [3]. The generating function in [5] is found by first deriving the probability mass function; the end result differs significantly from (23) because they ignore a critical detail, namely the dependence between the number of backoff slots of a node and the delay due to transmissions and collisions of competing stations. In a companion paper [13], we explain the differences in detail, and show how the analysis in [5] can be corrected to obtain (23). The generating function in [3] is derived from the well-known Markov chain model of DCF [4] using a transformation technique. Their solution is similar in structure to (23), but omits the normalization term η and displays a very different expression for the term corresponding to $\widehat{Y}(z)$ in (22). The latter difference results because the authors of [3] make no allowance for the reduced contention effect. Instead, they assume there can be multiple interruptions to each backoff slot and that all non-tagged stations can source such interruptions. In Section V, we compare the accuracy of (23) with that of the solutions in [5] and [3].

IV. Asymptotic analysis

In this section, we investigate the asymptotic behaviour predicted by our delay model. First, we consider the rate of growth of the mean access delay when the number of nodes $n \to \infty$. To obtain meaningful results under this regime, it must also be assumed that $m = \infty$ and $K = \infty$. The reason for this can be deduced from (4) and (1). For finite m or K, the attempt probability τ is always bounded away from zero, which means that $p \to 1$ as $n \to \infty$. The result is that no packets will be successful and the access delay will be undefined. Conversely, it has been shown in [11] and [12] that when $m = \infty$ and $K = \infty$,

$$\lim_{n\to\infty} \tau = 0, \tag{24}$$

$$\lim_{n\to\infty} n\tau \uparrow \ln(\frac{\lambda}{\lambda-1}), \tag{25}$$

$$\lim_{n\to\infty} p \uparrow 1/\lambda. \tag{26}$$

The following lemma states that $E[D] \sim O(n)$.

Lemma 1: For $m = \infty$ and $K = \infty$,

$$\lim_{n\to\infty} E[D] = n\left(\frac{(\lambda t_{slot} + E[C^*])}{\ln(\frac{\lambda}{\lambda-1})(\lambda-1)} + E[T^*] - E[C^*]\right). \tag{27}$$

Proof: It can be shown from (18) that when $m = \infty$ and $K = \infty$,

$$E[D] = \frac{(t_{slot} + pE[C^*])}{\tau(1-p)} + \frac{(n-1)(E[T^*] - E[C^*])}{1-\tau} + \frac{pE[C]}{(1-p)} + E[T].$$

Taking the limit and applying (24)-(26) leads to the result. ■

We see that the asymptotic mean delay depends on the backoff multiplier λ and the durations of collisions and transmissions but is independent of the initial window size W. Kwak et al. [11] previously obtained the $O(n)$ result for the asymptotic mean access delay, but they performed a simplified analysis of the delay which took only the backoff slots into account and ignored other contributions to the delay.

The above result naturally raises the following questions. How do higher order moments behave as n (and consequently p) is varied? What is the tail behaviour of the delay distribution? The next theorem sheds some light on both of these

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

questions. The theorem refers to the notion of a heavy-tailed distribution; we understand that a distribution F is heavy-tailed if its moment generating function diverges, namely $\int_0^\infty e^{\epsilon x} dF(x) = \infty \quad \forall\, \epsilon > 0$ [14].

Theorem 1: For $m = \infty$ and $K = \infty$,

i) the kth moment of the access delay D is finite for $0 \leqslant p < \frac{1}{\lambda^k}$, and infinite elsewhere,

ii) the access delay D has a heavy-tailed distribution.

Proof: From (13), we have $E[D^k] = E[(A+T)^k]$. For realistic traffic the r.v. T is bounded, implying that all its moments exist. In contrast, we will find that not all moments of A exist. We focus on the component of $E[D^k]$ which preserves the highest order moment in A, namely $E[A^k]$ given by

$$E[A^k] = (1-p)\sum_{i=0}^{\infty} p^i \,\mathrm{E}[(A^{(i)})^k]. \tag{28}$$

Again, since the r.v.'s C_{ij} are bounded, we focus on the component of (28) given by

$$(1-p)\sum_{i=0}^{\infty} p^i \,\mathrm{E}[(\sum_{j=0}^{i} B^{(j)})^k] = \sum_{i=0}^{\infty} p^i \,\mathrm{E}[(B^{(i)})^k] + \text{other terms}. \tag{29}$$

The other terms in (29) have moment order less than k so we ignore them. We consider the term $\mathrm{E}[(B^{(i)})^k]$ in the sum. It can be shown (e.g. using transforms) that

$$\mathrm{E}[(B^{(i)})^k] = E[(U^{(i)})^k](E[t_{slot} + Y])^k + \text{other terms},$$

where the other terms involve moments of $U^{(i)}$ of order less than k. Now, it can be shown that for a discrete uniform distribution on the integers $0, 1, \ldots, N-1$, the dominant term in the kth moment has the form $c_1 N^k$, for some constant c_1. Hence the sum $\sum_{i=0}^{\infty} p^i \,\mathrm{E}[(B^{(i)})^k]$ in (29) will contain the term

$$c_2 W^k \sum_{i=0}^{\infty} p^i \lambda^{ik} = \frac{c_2 W^k}{(1 - p\lambda^k)}, \tag{30}$$

for some constant c_2. Clearly, the sum (30) is finite for $0 \leqslant p < \lambda^{-k}$ but divergent elsewhere. The sum blows up due to the rapid (exponential) growth in the kth moment over the sequence of uniform distributions. All other terms that we have ignored in the steps of the derivation above involve lesser moments of the uniform distributions, and it is easy to see that these terms are finite for $0 \leqslant p < \lambda^{-k} + \epsilon$ for some $\epsilon > 0$. This completes the proof of (i).

Since there are infinite moments, the proof of (ii) is immediate. ∎

The theorem shows that for $\lambda > 1$, the number of bounded moments decreases as p increases. The proof of the theorem shows that the heavy tail is a direct consequence of the exponential growth of the backoff window in the EB process. In reality, of course, truncated BEB is implemented and the delay is always bounded. Nevertheless, Theorem 1 suggests that the access delay statistics for truncated BEB will contain precursors of heavy-tailed characteristics, such as large variance. We confirm this in Section V by demonstrating that our model yields a truncated heavy-tail for even moderate m and K.

TABLE I
802.11B MAC AND PHY PARAMETERS

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}	224 bits
UDP/IP header	l_{udpip}	320 bits
ACK packet	l_{ack}	112 bits
Slot time	t_{slot}	20 μs
SIFS	t_{sifs}	10 μs
DIFS	t_{difs}	50 μs
Min CW	W	32
Doubling limit	m	5
Retry limit	K	7

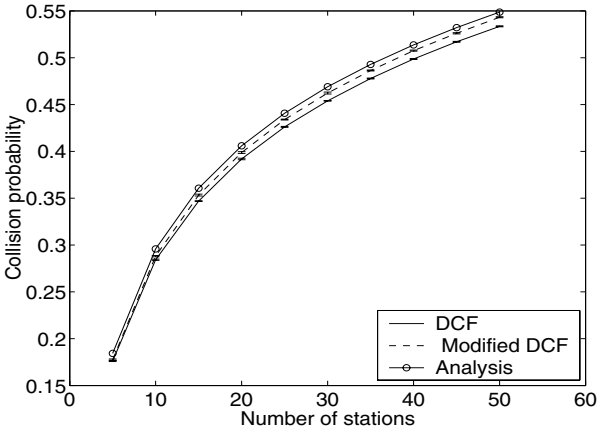


Fig. 1. Collision probability.

V. Numerical Evaluation and Discussion

The objectives of this section are to verify our analytical model with simulation, and to study the characteristics of the access delay as a function of n, m, K and the packet size. The simulations were performed using the ns-2 simulator [7] (version 2.27) which has a built-in implementation of the IEEE 802.11 MAC. Detailed testing revealed that the simulator contains several points of non-compliance with the IEEE 802.11 MAC standard [8] that noticeably affect the output delay statistics, so these were remedied. The main problems with the standard simulator are: the timer modelling the DIFS deferral is not stopped when the channel becomes busy; a post-backoff is not preceded by a DIFS; after the backoff counter is frozen, the remaining backoff time is incorrectly calculated; and the EIFS period is erroneously followed by a DIFS deferral.

We simulate a network scenario comprised of n saturated stations sending packets to an access point, in ideal channel conditions. The stations use the UDP protocol with a fixed packet size. We choose MAC and physical layer parameter values consistent with an 802.11b system [15]. Table I lists the parameters, the symbols that we use for them, and their values. Denoting the UDP packet payload by l_{pay} bits, the packet transmission times used in our analytic models are given by $t_{data} = t_{phys} + \frac{l_{mac}+l_{udpip}+l_{pay}}{r_{data}}$, and $t_{ack} = t_{phys} + \frac{l_{ack}}{r_{ctrl}}$. We ignore propagation delays since they are several orders of magnitude smaller.

First we examine the agreement between the collision

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

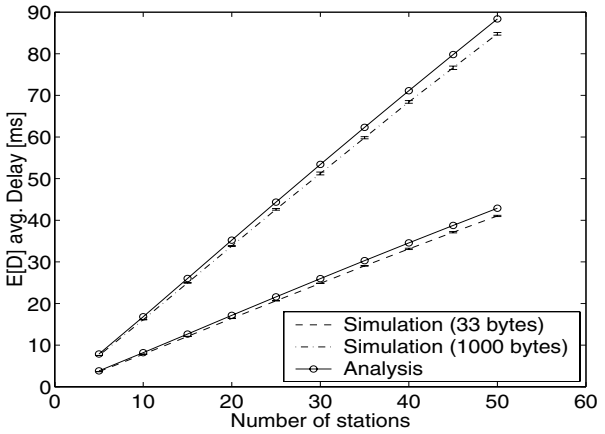


Fig. 2. Average access delay for l_{pay} =33 and 1000 bytes.

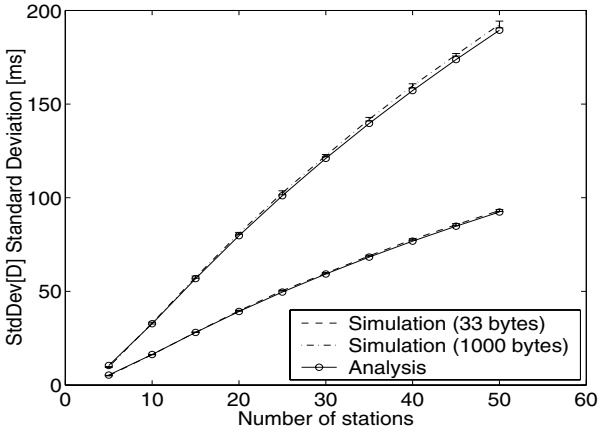


Fig. 3. Standard deviation of access delay for l_{pay} =33 and 1000 bytes.

probability measured from simulation and that calculated using (1) and (4). Fig. 1 displays the collision probability as a function of the number of stations n, where the simulation curve is labeled 'DCF'. The simulation results are shown with 95% confidence intervals. Observe that the fixed point approximation provides a reasonable estimate for the collision probability, although the approximation becomes less accurate as n increases. The discrepancy is partly due to the reduced contention effect, which is not addressed by the analytical model. To show this, we eliminated the reduced contention effect in the ns-2 simulator by modifying the backoff operation to decrement at the end of every DIFS as well as at the end of every idle slot. The results, labeled 'modified DCF' in Fig. 1, indeed give a simulation curve that is closer to the analytical result.

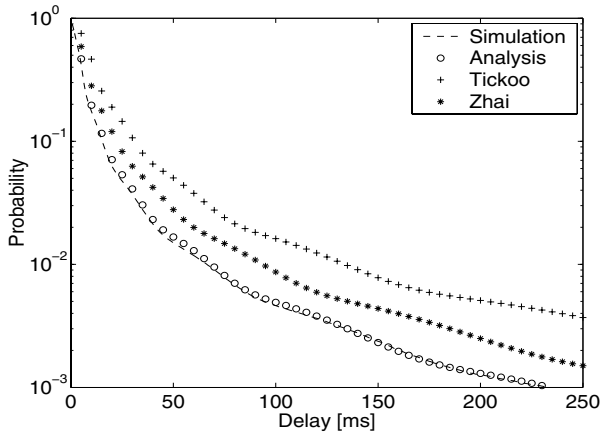
In Fig. 2 we plot the average access delay of DCF as a function of n using both simulation and our analytic formula (18), for UDP payloads l_{pay} =33 and 1000 bytes. We give results for up to $n = 50$ active nodes, which we believe encompasses the size of active user populations encountered in most IEEE 802.11b deployments. We plot the corresponding results for the standard deviation of the delay in Fig. 3, where the analytic curve is obtained by taking the square root of (19). Although calculation of the delay moments relies on an approximate p that ignores the reduced contention effect, the results obtained from our analysis match our simulation results very well. From the graphs we conclude that the accuracy of our analytical model is also maintained for a range of different packet sizes. Observe that for the plotted range of n, the mean delay and jitter each station experiences increases dramatically with n, which is due to each node experiencing more collisions and more interruptions to its backoff timer. The durations of collisions and interruptions also has an impact, as evidenced by the increased delays for the larger packets.

Next, we compare the CCDF of the access delay obtained by numerically inverting (20) with that obtained from simulation. The LATTICE-POISSON inversion algorithm developed by Abate, Choudhury and Whitt [6] was used, with parameters selected to give an inversion error no greater than 10^{-8} and the lattice spacing $\delta = 10\mu s$. In Figs. 4(a) and 4(b) we plot results for $n = 10$ and l_{pay} =33 and 1000 bytes, respectively. Figs. 4(c) and 4(d) display results for $n = 30$ and l_{pay} =33 and 1000 bytes, respectively. All results confirm that our model is very accurate, even for small tail probabilities. The analytical points do not merely predict the general trend of the simulation curves, they actually follow most of the undulations in the curves. In contrast, inverting the generating functions derived by Tickoo and Sikdar [5], and Zhai, Kwon and Fang [3] leads to distribution results which are far from the simulation curves.

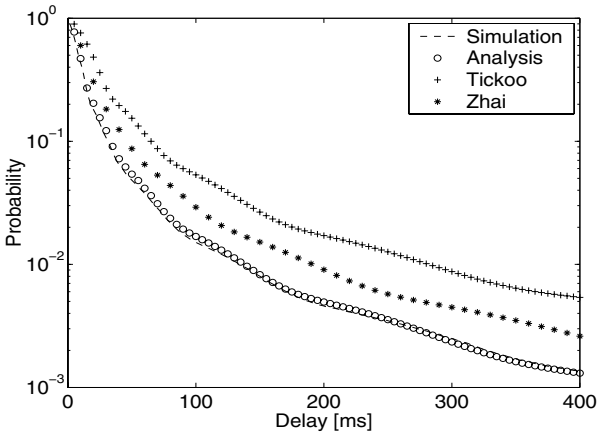
Finally, we investigate what the asymptotic results can tell us about practical operating regimes. Recall that the asymptotic results are for infinite m and K so there is no discarding of packets. Therefore, we expect that the asymptotic results can only inform about protocol performance for finite m and K for loading regimes that do not result in high rates of packet discards. In Fig. 5, we plot for $0 < n \leq 300$ the linear asymptotic result for mean delay (27), as well as the analytic mean (18) for various pairs of m and K. We see that the analytic curves agree quite well with the asymptote for relatively small n, but they fall away as n is increased due to more packets reaching the m and K limits. When m and K are increased, the agreement is maintained for a wider range of n. In Fig. 6, we plot the analytic standard deviation (19) for various m and K pairs. Theorem 1 tells us that for infinite m and K, the standard deviation grows rapidly with load and is ultimately unbounded. Fig. 6 shows an initial rapid growth of the standard deviation as n increases, with the rate of growth increasing with m and K, before hitting a plateau due to the m and K limits. Fig. 7 shows the analytic CCDF (20) for n =10 and 30 and various m and K pairs plotted on a log-log scale. We observe the precursors of heavy-tailedness suggested by Theorem 1. There is an initial linearity followed by a faster roll-off, which are characteristics of a truncated power-law tail distribution. The power-law exponent decreases with increasing n and the point of truncation moves further into the tail with increasing m and K.

Before closing this section, we comment on the relevance of our findings on heavy tails to real wireless LANs. While our analysis has been for the saturated setting, we argue that episodes of long delays and large delay variance are possible for statistical traffic if there are concurrent long bursts or a sufficient number of stations. Indeed, empirical evidence for extremely long access delays due to truncated BEB have

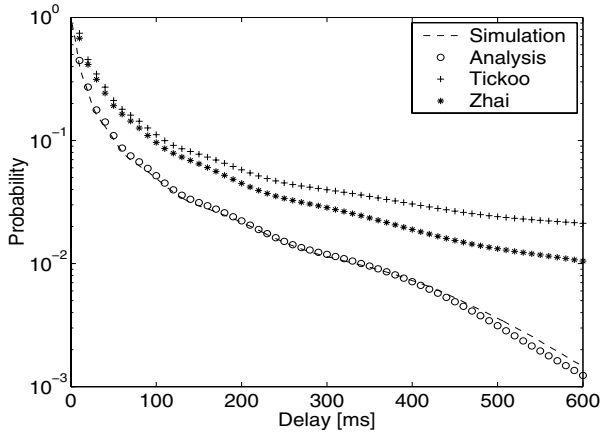
Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.



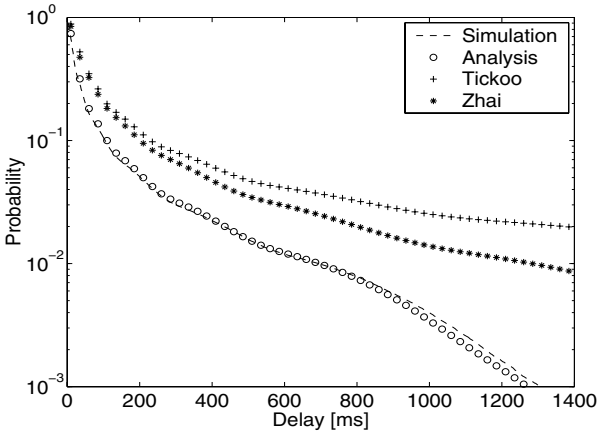
(a) $n = 10$ and $l_{pay} = 33$ bytes.



(b) $n = 10$ and $l_{pay} = 1000$ bytes.



(c) $n = 30$ and $l_{pay} = 33$ bytes.



(d) $n = 30$ and $l_{pay} = 1000$ bytes.

Fig. 4. Complementary cumulative distribution function (CCDF) of access delay.

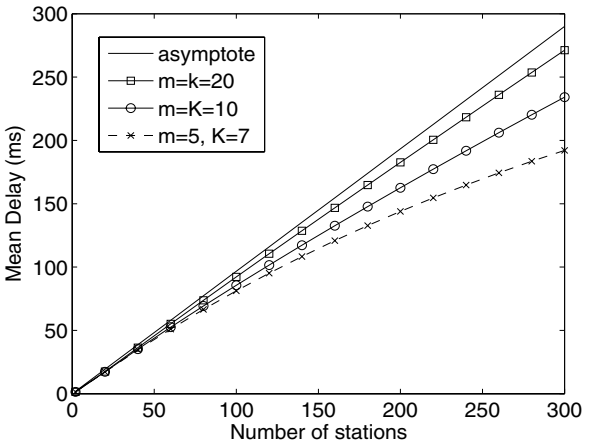


Fig. 5. Analytical mean delay for $l_{pay} = 33$ bytes and different m and K, compared with asymptote.

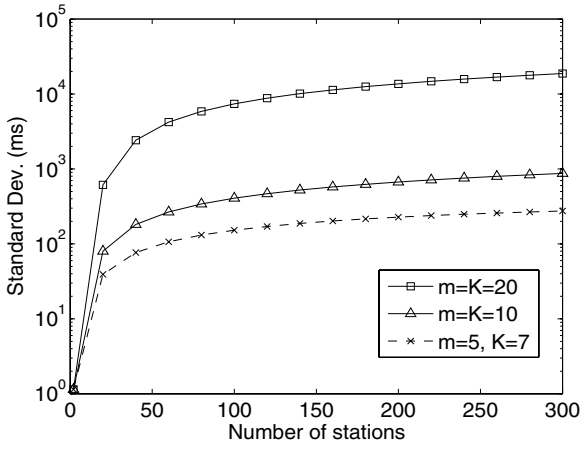


Fig. 6. Analytical standard deviation of delay for $l_{pay} = 33$ bytes and different m and K.

been reported in measurements on Ethernet [16]. A heavy-tailed access delay leads to an even heavier queueing delay distribution. Large delays not only degrade the quality of real-time services, but if delays are extremely long, there is the possibility of causing timeouts in non-real-time higher layer protocols such as TCP, resulting in throughput degradation.

VI. Conclusion

In this paper, we have developed a model of the access delay of the IEEE 802.11 MAC for saturated stations. The model enables a unified analysis of the moments and generating function. We have shown how numerical transform inversion can be used to compute distributional values from the generating

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.

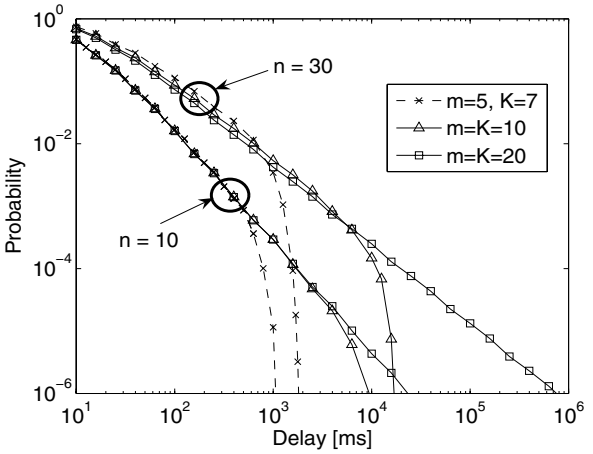


Fig. 7. Analytical CCDF of delay for $l_{pay} = 1000$ bytes and different n, m and K.

function. Our formulae for the mean, standard deviation and generating function are much more accurate than others that have appeared in the literature [2], [5], [3].

Our asymptotic analysis of the theoretical unlimited retransmission case provides insights for practical DCF systems. In particular, we have shown that the heavy-tail induced by BEB in the theoretical system translates to a truncated power-law tail induced by truncated BEB in DCF. This result implies a relatively high probability of long packet delays in DCF and raises doubts about the efficacy of using DCF for delay-sensitive applications.

REFERENCES

[1] J. Håstad, T. Leighton, and B. Rogoff, "Analysis of backoff protocols for multiple access channels," *SIAM J. Comput.*, vol. 25, pp. 740–774, 1996.

[2] M. M. Carvalho and J. J. Garcia-Luna-Aceves, "Delay analysis of IEEE 802.11 in single-hop networks," in *Proc. of 11th IEEE International Conference on Network Protocols (ICNP)*, Atlanta, 2003, pp. 146–155.

[3] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *Wireless Commun. Mobile Comput.*, vol. 4, pp. 917–931, 2004.

[4] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, pp. 535–547, 2000.

[5] O. Tickoo and B. Sikdar, "Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks," in *Proc. IEEE INFOCOM 2004*, pp. 1404–1413.

[6] J. Abate, G. L. Choudhury, and W. Whitt, "An introduction to numerical transform inversion and its application to probability models," in *Computational Probability*, W. K. Grassman, Ed. Norwell, MA: Kluwer, 2000, pp. 257–323.

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

[8] IEEE, *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, 1999.

[9] 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, July/Aug. 2005.

[10] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *Proc. IEEE INFOCOM 2002*, pp. 599–607.

[11] B.-J. Kwak, N.-O. Song, and L. E. Miller, "Performance analysis of exponential backoff," *IEEE/ACM Trans. Networking*, vol. 13, pp. 343–355, 2005.

[12] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New insights from a fixed point analysis of single cell IEEE 802.11 WLANs," in *Proc. IEEE INFOCOM 2005*, pp. 1550–1561.

[13] H. L. Vu and T. Sakurai, "Accurate delay distribution for IEEE 802.11 DCF," *IEEE Commun. Lett.*, vol. 10, pp. 317–319, 2006.

[14] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels, *Stochastic Processes for Insurance and Finance*. Chichester: Wiley, 1999.

[15] IEEE, *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher-Speed Physical Layer (PHY) Extension in the 2.4 GHz Band*, 1999.

[16] M. Molle, "A new binary logarithmic arbitration mechanism for Ethernet," Computer Systems Research Institute, University of Toronto, Canada," Technical Report CSRI-298, 1994.

Taka Sakurai (M'02) received the B.Sc. degree in applied mathematics and B.E. degree in electrical engineering from the University of Adelaide in 1988 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Melbourne in 2003. From 1991 to 1997, he was a Research Engineer at Telstra Research Laboratories. Subsequently, he held research and development roles at NEC and Lucent Technologies. From 2003 to 2005, he was with the Department of Electrical and Electronic Engineering, University of Melbourne. Currently, he is with the Chief Technology Office of Telstra Corporation, and an Honorary Fellow of the University of Melbourne. His research interests are in the areas of performance analysis of wireless networks, design of MAC protocols for wireless LANS and sensor networks, and computational probability.

Hai L. Vu (S'97-M'98-SM'06) received the B.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 he is now with the Centre for Advanced Internet Architectures (CAIA), Faculty of Information and Communication Technologies (FICT), Swinburne University of Technology, Hawthorn, Victoria, Australia.

He has authored or coauthored over 50 scientific journals and conference papers. His current research interests are in data network modeling, performance evaluation of wireless and optical networks, and telecommunication networks design.

Authorized licensed use limited to: SWINBURNE UNIV OF TECHNOLOGY. Downloaded on February 15, 2009 at 17:57 from IEEE Xplore. Restrictions apply.