An Access Delay Model for IEEE 802.11e EDCA

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

Abstract—We analyze the MAC access delay of the IEEE 802.11e enhanced distributed channel access (EDCA) mechanism under saturation. We develop a detailed analytical model to evaluate the influence of all EDCA differentiation parameters, namely AIFS, CWmin, CWmax, and TXOP limit, as well as the backoff multiplier β. Explicit expressions for the mean, standard deviation, and generating function of the access delay distribution are derived. By applying numerical inversion on the generating function, we are able to efficiently compute values of the distribution. Comparison with simulation confirms the accuracy of our analytical model over a wide range of operating conditions. We derive simple asymptotics and approximations for the mean and standard deviation of the access delay, which reveal the salient model parameters for performance under different differentiation mechanisms. We also use the model to numerically study the differentiation performance and find that β differentiation, though rejected during the standardization process, is an effective differentiation mechanism that has some advantages over the other mechanisms.

Index Terms—Medium access delay, IEEE 802.11e, QoS, EDCA, service differentiation, generating function.

1 Introduction

A quality-of-service (QoS) extension to the original IEEE 802.11 wireless local area network standard [1], known as IEEE 802.11e [2], defines a contention-based medium access control (MAC) scheme called *enhanced distributed channel access* (EDCA). EDCA provides service differentiation by separating flows into different access classes. The differentiation achieved by EDCA is relatively easy to understand in a qualitative sense; however, quantifying the degree of differentiation provided is difficult due to the distributed, contention-based nature of EDCA. Hence, there is a need for accurate performance models to guide the configuration of parameters. In this paper, we develop a detailed analytical model of the packet access delay in a network of 802.11e EDCA stations operating under saturation. In this context, access delay is the time interval between the instant a packet reaches the head of the transmission queue, and the time when the packet is successfully received at the destination station.

Service differentiation in EDCA is effected through four parameterized access categories (ACs). Packets belonging to different ACs are given different access priorities by appropriate tuning of four AC-specific parameters. The parameters define, respectively, the size of AC-dependent guard periods (arbitrary interframe spacing or AIFS), minimum and maximum contention windows (CWmin and CWmax), and lengths of packet bursts or transmission opportunity limit (TXOP limit). A fifth parameter representing the backoff window multiplier, which we denote by β, was studied during the standardization process, but was eventually abandoned due to doubts about effectiveness [3] and replaced with a fixed multiplier of 2. In this paper, we substantially extend a model [4] that we developed previously for access delay in the distributed coordination function (DCF) of the original IEEE 802.11 MAC, to EDCA. Our model[1] can scale to an arbitrary number of ACs and accounts for all four standardized differentiation parameters. We also make our model general enough to cover β differentiation, so that we can study the characteristics of this mechanism.

Many recent papers have proposed analytical models for various subsets of EDCA functionality [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Xiao [7] models CWmin and CWmax differentiation, the authors of [8], [9], [10], [11], [12], [13], [14], [15], and [16] model CWmin, CWmax, and AIFS differentiation, and Peng et al. [17] develop a simple model for TXOP differentiation only. Compared to previous models, our model is novel for the following reasons: 1) it correctly accounts for all four differentiation parameters in the standard, 2) it yields the standard deviation and distributional values of the access delay, as well as the commonly obtained mean access delay, and 3) it provides accurate estimates of these metrics. Ge at al. [18] attempt to explicitly account for all differentiation parameters in their model, but they actually analyze and simulate a p-persistent version of EDCA, which does not have the same characteristics as

- *D. Xu is with the National ICT of Australia (NICTA), Victoria Laboratory, Department of Electrical and Electronic Engineering, University of Melbourne, Building 193, Melbourne VIC 3010, Australia. E-mail: d.xu@ee.unimelb.edu.au.*
- *T. Sakurai is with the Chief Technology Office, Telstra Corp., 13/242 Exhibition St., Melbourne VIC 3000, and also with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne VIC 3010, Australia. E-mail: tsakurai@ieee.org.*
- *H.L. Vu is with the Centre for Advanced Internet Architectures (CAIA), EN Building, Room 606b, Mail H39, Faculty of Information and Communication Technologies, Swinburne University of Technology, Hawthorn, Melbourne VIC 3122, Australia. E-mail: hvu@swin.edu.au.*

Manuscript received 1 May 2007; revised 31 Jan. 2008; accepted 7 July 2008; published online 17 July 2008.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2007-05-0123. Digital Object Identifier no. 10.1109/TMC.2008.108.

1. Parts of this work have appeared previously in conference form [5], [6].

EDCA. In [12], it is stated that a four-parameter model can be built by simply inflating the packet length in their three-parameter model to account for TXOP differentiation. However, as we will show in our model development, an accurate model of TXOP differentiation is a nontrivial extension that requires careful consideration of all possible combinations of transmission and collision durations of the different ACs, together with their probabilities of occurrence.

Our analytical model is fully integrated and can capture joint differentiation by up to four parameters (or five including β). However, for ease of understanding, we present the model as three submodels: a collision probability model that estimates the collision probabilities of the different classes, a delay model that accounts for all events that contribute to the access delay, and a TXOP model that accounts for TXOP differentiation. The collision probability and delay models capture the influence of the CWmin, CWmax, and AIFS mechanisms. By virtue of the way in which the TXOP mechanism operates, it becomes natural to treat it as a modeling extension.

A collision probability model is a vital element of any EDCA analysis. All the aforementioned studies use extensions of Bianchi's 2D Markov chain analysis of DCF [19] to derive the collision probabilities. To incorporate AIFS differentiation, the authors of [8] and [9] resort to 3D Markov chains, while Tsai and Wu [10] use a 4D Markov chain. In contrast, the authors of [14] and [15] develop less complex models based on separate 2D and 1D Markov chains. Our collision probability model is based on that of [14], but uses an average value analysis in place of the 2D Markov chain. This leads to a more intuitive and simple, yet accurate collision probability model.

Our delay and TXOP models are novel and yield detailed statistics of the access delay. Most prior studies of EDCA analyze only throughput and/or mean delay. Exceptions are [9], where the delay distribution is calculated based on the transient analysis of a Markov chain; [16], where the delay distribution is approximated by estimating the probabilities of alternate delay outcomes; and [13], where points of the distribution are obtained by inverting the generating function of the delay distribution. We present a more direct and accurate method to obtain the distribution. Similar to [13], we derive the generating function of the access delay distribution and obtain distributional values via numerical transform inversion. However, our generating function is more accurate than that of [13], as we illustrate through a numerical comparison. Further, we obtain explicit expressions for the mean and standard deviation of the access delay. Our moment expressions are derived via direct probabilistic arguments, in contrast to [13], where differentiation and limit taking of the generating function is advocated. The direct approach is easier because the generating function in question is complicated, making differentiation tedious. Perhaps as a concession to this complexity, Engelstad and Østerbø [13] go no further than state the standard deviation in terms of derivatives of the generating function. As far as we are aware, ours is the first work to obtain an explicit expression for the standard deviation of the delay (or jitter) in EDCA. The expression enables us to develop analytical insights into the relative importance of parameters and to quantify the jitter performance of the differentiation mechanisms.

We confirm that our analytical results for the mean, standard deviation, and distribution of the access delay are accurate through comparison with ns-2 simulation. Typically, our analytical tail distribution is an excellent match with simulation down to 10^{-3}, and often beyond.

In addition to developing a model, we exploit the model to advance the understanding of EDCA delay performance. We use the model to derive *asymptotics* for the mean under the assumptions of unlimited retransmissions and the number of stations tending to infinity, and to derive *approximations* for both the mean and standard deviation under the assumptions of a finite retransmission limit and a large number of stations. The asymptotics and approximations reveal the salient model parameters for performance under each differentiation mechanism, and provide simpler alternatives to the complete analytical expressions for system analysis and design. Our approximation methodology and results are new. Our asymptotic work is inspired by that of Ramaiyan et al. [20], who obtained asymptotics for throughput ratios under CWmin, AIFS, and β differentiation. There are some parallels between their asymptotic throughput ratios and our asymptotic mean delay ratios (since under infinite retransmissions, the mean access delay has a simple relationship with the throughput). Unlike [20], we derive asymptotic results for the individual ACs as well as the ratios, and a result for TXOP differentiation.

Finally, we perform a detailed numerical study using the analytical model to quantify the differentiation in the mean and standard deviation afforded by CWmin, AIFS, TXOP limit, and β. We find that β differentiation, though discarded during the standardization process, is an effective differentiation mechanism that has some advantages over the other mechanisms.

The rest of this paper is organized as follows: In Section 2, we present our analytical model, starting with the collision probability model. Next, we describe our access delay model that accounts for the CWmin, CWmax, AIFS, and β mechanisms, and derive expressions for the associated moments and generating function. At the end of this section, we present our TXOP model, and derive the moments and generating function when all five differentiation parameters are included. In Section 3, we present asymptotics and approximations. Validation of the analytical model with ns-2 simulation is carried out in Section 4, and then we use the model to assess the nature of the service separation provided by each differentiation mechanism, and to test the accuracy of the approximations. Finally, we state our conclusions in Section 5. Some details not provided in this paper, such as further explanations of derivations and additional numerical results, can be found in the technical report [21].

2 Analytical Model

EDCA realizes service differentiation through the use of four parameterized ACs in each station. The CWmin and CWmax parameters define the initial and maximum values of the contention window (CW) used in the backoff process, whereby a discrete backoff time measured in backoff slots

is randomly selected from [0, CW-1]. The AIFS parameter defines the guard time that a station must observe after a busy channel period. A smaller AIFS means a higher priority of access. The TXOP limit parameter defines the maximum duration for which a station can enjoy uninterrupted control of the medium after obtaining a transmission opportunity. A value of $\mathrm{TXOP\ limit} = 0$ indicates only a single packet may be transmitted for each transmission opportunity. EDCA can operate in either two-way (DATA-ACK) or four-way (RTS-CTS-DATA-ACK) handshaking modes. Our analysis covers the former, but can be readily extended to the latter.

In our model, we make the following assumptions:

1. All stations are saturated (always have a packet to send).
2. The collision probability is constant regardless of the state, but may differ with AC.
3. Channel conditions are ideal.
4. ACK packets are transmitted at the lowest basic rate, and the ACK timeout after a collision matches the guard time observed by noncolliding nodes.
5. Each station only has traffic belonging to a single AC.

The first four assumptions are common for studies of 802.11 performance and originate from [19]. Assumptions 4 and 5 can be removed at the expense of additional modeling complexity.

We allow for an arbitrary J distinct ACs in the network. Without loss of generality, we label the ACs with indices $k = 1, \ldots, J$, in order of nondecreasing AIFS, while placing no ordering restrictions on the values of the other AC parameters. We refer to the kth AC as $\mathrm{AC}[k]$, and denote the associated AIFS period by AIFS_k. The number of $\mathrm{AC}[k]$ stations is denoted by n_k, R is the maximum number of attempts (the same for all ACs as specified in [2]), and W_k is the minimum contention window for $\mathrm{AC}[k]$. We generalize the backoff mechanism in this paper to exponential backoff with real multiplier $\beta_k > 1$. The maximum backoff stage for $\mathrm{AC}[k]$ is m_k, so that the maximum contention window, denoted by M_k, is obtained by $\langle \beta_k^{m_k} \mathrm{W}_k \rangle$, where $\langle . \rangle$ denotes rounding to the nearest integer. The transmission opportunity limit for $\mathrm{AC}[k]$ is denoted by TXOP_k.

2.1 Collision Probability Model

Our objective is to develop a fixed-point approximation to compute the collision probabilities and transmission probabilities of all the ACs. Let c_k and p_k denote the collision probability and transmission probability, respectively, experienced by an $\mathrm{AC}[k]$ packet. The fixed-point approximation is established by combining a set of equations for the c_k's expressed in terms of the p_k's, with an opposing set of equations for the p_k's expressed in terms of the c_k's. We obtain the former set of equations by following an approach proposed in [14], summarized below.

In [14], and also in [15], the concept of *slot class* is used to account for the effect of AIFS differentiation on the collision probability. Slot class can be understood with the aid of Fig. 1, where we illustrate a particular configuration of AIFS_k parameters. Let us number the idle slots after an AIFS_1 with *slot numbers*, starting from 1. The increase in the AIFS_k values with k restricts the slots in which higher-numbered ACs can compete for channel access. For

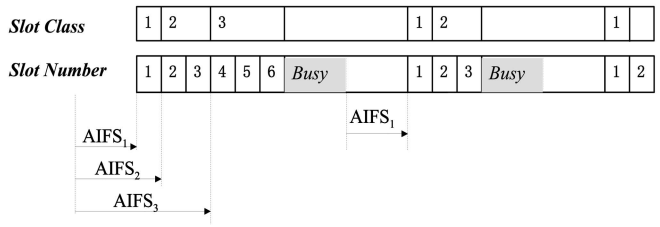

Fig. 1. Slot number and slot class.

example, while $\mathrm{AC}[1]$ stations can begin to compete for the channel access in slot number 1, $\mathrm{AC}[2]$ stations can only begin from slot number 2. In line with this observation, we divide the slots into numbered groups called slot classes, where the slot class number corresponds to that of the highest numbered AC that may compete for access.

In slot class j, only stations with AC $k \leq j$ can transmit. This gives rise to the notion of a conditional collision probability $c_k(j)$ for $\mathrm{AC}[k]$ in slot class j, given by

$$c_k(j) = 1 - \frac{\prod_{i=1}^{j} r_i^{n_i}}{r_k}, \ (k \leq j), \tag{1}$$

where we denote $r_i = 1 - p_i$, $i = 1, \ldots, J$.

The overall collision probability c_k is obtained as an average of the $c_k(j)$'s weighted by the stationary probabilities $P(j)$ that a randomly selected slot belongs to slot class j:

$$c_k = \sum_{j=k}^{J} c_k(j) \frac{P(j)}{\sum_{i=k}^{J} P(i)}. \tag{2}$$

The probabilities $P(j)$ can be found by examining the evolution of the slot number/class. In [14], it is shown that the evolution can be described by a Markov chain. Each state of the Markov chain represents a slot number, and a transition is made at each slot according to whether the slot is idle or marks the beginning of a successful transmission or collision. If the slot is idle, the slot number is increased by one; if it is not idle, the slot number is reset to 1. The probabilities $P(j)$ can be computed from the steady-state probabilities of the Markov chain as

$$\begin{aligned} P(j) &= \frac{Q(j)}{\sum_{i=1}^{J} Q(i)}, \\ Q(j) &= \frac{1 - \alpha_j^{h^{(j+1)} - h^{(j)}}}{1 - \alpha_j} \prod_{i=1}^{j-1} \alpha_i^{h^{(i+1)} - h^{(i)}}, \end{aligned} \tag{3}$$

where we define $\prod_{i=1}^{0} \alpha_i^{h^{(i+1)} - h^{(i)}} = 1$, and

$$\alpha_j = \prod_{k=1}^{j} r_k^{n_k}, \qquad h^{(j)} = \frac{\mathrm{AIFS}_j - \mathrm{AIFS}_1}{t_{slot}}.$$

Equations (1), (2), and (3) express c_k as a nonlinear function of the transmission probabilities p_k. To find p_k as a function of the collision probabilities c_k, the authors of [14] and [15] use variants of the 2D Markov chain of [19]. In contrast, we invoke a mean-value approximation for p_k by equating it to the reciprocal of the *average backoff period* of an

AC[k] station. In other words, if Ψ_k is the average backoff period, then we write

$$p_k = \frac{1}{\Psi_k}. \tag{4}$$

To find the average backoff period, we analyze the dynamics of the backoff process in a similar way to Kwak et al. [22], who analyzed the backoff process for DCF. The evolution of the backoff process of an AC[k] station at transmission instants can be described by a discrete-time Markov chain $s(t)$ with nonzero transition probabilities:

$$\begin{aligned}
&P(s(t+1) = i | s(t) = i-1) = c_k, \quad i = 1, \ldots, R-1,\\
&P(s(t+1) = 0 | s(t) = i) = 1 - c_k, \quad i = 0, \ldots, R-2,\\
&P(s(t+1) = 0 | s(t) = R-1) = 1, \quad i = R-1.
\end{aligned}$$

It is straightforward to show that the steady-state probabilities of $s(t)$ are given by $\pi_i^{(k)} = (1-c_k)c_k^i(1-c_k^R)^{-1}$, for $i = 0, \ldots, R-1$.

Let $U_i^{(k)}$ be a discrete uniform random variable (r.v.) representing the backoff duration that an AC[k] station has to wait in the ith backoff stage. These r.v.'s have densities defined by

$$P\left[U_i^{(k)} = j\right] = \begin{cases} u\left(0, \langle \beta_k^i \mathrm{W}_k \rangle - 1\right), & \text{for } i = 0, \ldots, m_k - 1,\\ u\left(0, \langle \beta_k^{m_k} \mathrm{W}_k \rangle - 1\right), & \text{for } i = m_k, \ldots, R-1, \end{cases} \tag{5}$$

where $u(a,b)$ is the discrete uniform density with support $(a, \ldots, b)$. The corresponding average backoff durations, $\mathrm{E}[U_i^{(k)}]$, are given by

$$\mathrm{E}\left[U_i^{(k)}\right] = \begin{cases} \frac{\langle \beta_k^i \mathrm{W}_k \rangle - 1}{2}, & \text{for } i = 0, \ldots, m_k - 1,\\ \frac{\langle \beta_k^{m_k} \mathrm{W}_k \rangle - 1}{2}, & \text{for } i = m_k, \ldots, R-1. \end{cases} \tag{6}$$

Knowing the steady-state probabilities and average durations of the R backoff stages, it follows that the overall average backoff period of an AC[k] station is

$$\begin{aligned}
\Psi_k &= \sum_{i=0}^{R-1} \pi_i^{(k)} \mathrm{E}\left[U_i^{(k)}\right]\\
&= \sum_{i=0}^{m_k-1} \eta_k c_k^i \left(\frac{\langle \beta_k^i W_k \rangle - 1}{2}\right)\\
&\quad + \sum_{i=m_k}^{R-1} \eta_k c_k^i \left(\frac{\langle \beta_k^{m_k} W_k \rangle - 1}{2}\right),
\end{aligned} \tag{7}$$

where $\eta_k = (1-c_k)(1-c_k^R)^{-1}$. Equations (1), (2), (3), (4), and (7) constitute a nonlinear system of equations that can be solved iteratively to obtain the p_k's and c_k's.

2.2 Delay Model

We consider a selected (tagged) AC[k] station and derive an expression for the access delay as experienced by packets of this station under saturation. Several events that contribute to the access delay can be identified, the most obvious being the successful transmission of the packet. Preceding this event will be the first backoff plus a variable number of collisions involving the tagged station and the associated 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 $D^{(k)}$ of the tagged station is

$$D^{(k)} = \epsilon^{(k)} + A^{(k)} + T^{(k)}, \tag{8}$$

where $\epsilon^{(k)}$ is an r.v. representing a defer period, which includes the duration of AIFS_k and the interruptions to this duration from higher priority stations; $A^{(k)}$ is an r.v. representing the sum of the durations of backoffs and collisions involving the tagged station, as well as the durations of successful transmissions and collisions of nontagged stations that interrupt the backoff timer of the tagged station. The last term, $T^{(k)}$, is the transmission time of the packet by the tagged station.

As mentioned previously, we first focus on the case of $\mathrm{TXOP}_i = 0$ $(i = 1, \ldots, J)$, which means only one packet transmission is permitted per channel access. In the case of fixed length data packets, this means that $T^{(k)} = t_{data}$, where t_{data} denotes the transmission time of a single data packet. In Section 2.5, we will remove this restriction on TXOP_i.

The defer period $\epsilon^{(k)}$ accounts for the duration of AIFS_k, as well as any interruptions to AIFS_k by transmissions from higher priority stations, namely AC[j] stations where $j < k$. Since $\mathrm{AIFS}_j < \mathrm{AIFS}_k$, an AC[$j$] station has the right to access the channel before the channel has been idle for AIFS_k. In this event, the tagged station resets the AIFS_k timer and starts a new countdown once the channel becomes idle again. Therefore, any number of interruptions by AC[j] stations are possible before AIFS_k can be successfully counted down.

We now obtain an expression for $\epsilon^{(k)}$. Clearly, $\epsilon^{(1)} = \mathrm{AIFS}_1$ since there is no interruption to the highest priority stations. On the other hand, the defer period for AC[k] stations with $k > 1$ must account for interruptions by any higher priority stations in any of the $h^{(k)}$ slots. As in Section 2.1, we refer to the successive idle slots following AIFS_1 as slots 1 to $h^{(k)}$. We denote $\varphi(i)$ as the slot class to which slot i belongs. The probability that at least one higher priority station transmits in slot 1 is

$$\mu_1 = 1 - \prod_{i=1}^{\varphi(1)} r_i^{n_i}. \tag{9}$$

The excess time due to an interruption in slot 1 from the point of view of the tagged station is

$$t_1 = \mathrm{AIFS}_1 + X_1. \tag{10}$$

The r.v. X_i represents the duration of the interruption in slot i; it could be a successful transmission when only one transmission occurs, or a collision when more than one station attempts to transmit.

If there is no transmission in slot 1, the probability that at least one higher priority station transmits in slot 2 is

$$\mu_2 = \prod_{i=1}^{\varphi(1)} r_i^{n_i} \left(1 - \prod_{j=1}^{\varphi(2)} r_j^{n_j}\right), \tag{11}$$

and the excess time for the tagged station is

$$t_2 = \mathrm{AIFS}_1 + t_{slot} + X_2. \tag{12}$$

This argument can be continued for all $h^{(k)}$ slots; the respective quantities for slot $h^{(k)}$ are

$$\begin{aligned} \mu_{h^{(k)}} &= \left[\prod_{i=1}^{h^{(k)}-1}\prod_{j=1}^{\varphi(i)} r_j^{n_j}\right]\left[1-\prod_{l=1}^{\varphi\left(h^{(k)}\right)} r_l^{n_l}\right], \\ t_{h^{(k)}} &= \mathrm{AIFS}_1 + \left(h^{(k)}-1\right)t_{slot} + X_{h^{(k)}}. \end{aligned} \tag{13}$$

The duration of interruptions X_i $(i = 1, \ldots, h^{(k)})$ can be expressed as

$$X_i = \begin{cases} T^*, & \text{w.p.} \quad \rho(i), \\ C^*, & \text{w.p.} \quad 1-\rho(i), \end{cases} \tag{14}$$

where w.p. stands for "with probability," T^* is the channel occupancy of a successful transmission from a higher priority station, and C^* is the channel occupancy of a collision involving higher priority stations. The quantity $\rho(i)$ is the probability of a successful transmission, conditional on at least one transmission. In the case when all data packets in the system are uniform and have fixed length, we have[2]

$$T^* = C^* = t_{data} + \mathrm{SIFS} + t_{ack}$$

and

$$\rho(i) = \frac{\sum_{l=1}^{\varphi(i)} n_l p_l r_l^{n_l-1}\prod_{j=1,j\neq l}^{\varphi(i)} r_j^{n_j}}{1-\prod_{l=1}^{\varphi(i)} r_l^{n_l}}. \tag{15}$$

The numerator in (15) is the probability of *exactly* one transmission, while the denominator is the probability of *at least* one transmission.

The defer period $\epsilon^{(k)}$ can be viewed as the waiting time until the first success in a sequence of independent trials, where each trial has $h^{(k)}+1$ possible outcomes corresponding to the $h^{(k)}$ types of interrupts plus the successful countdown of AIFS_k. The probability of a successful AIFS_k countdown is

$$s^{(k)} = 1 - \sum_{j=1}^{h^{(k)}} \mu_j = \prod_{i=1}^{h^{(k)}}\prod_{j=1}^{\varphi(i)} r_j^{n_j}. \tag{16}$$

Putting everything together, $\epsilon^{(k)}$ is given by

$$\begin{aligned} \epsilon^{(k)} &= i_1 t_1 + i_2 t_2 + \cdots + i_{h^{(k)}} t_{h^{(k)}} + \mathrm{AIFS}_k \\ &\text{w.p.} \quad \frac{\left(\sum_{l=1}^{h^{(k)}} i_l\right)!}{\prod_{l=1}^{h^{(k)}} i_l!}\mu_1^{i_1}\mu_2^{i_2}\cdots\mu_{h^{(k)}}^{i_{h^{(k)}}} s^{(k)}. \end{aligned} \tag{17}$$

The integers $i_1, i_2, \ldots, i_{h^{(k)}} = 0, 1, \ldots \infty$ represent the number of interruptions to each type of slot, and they extend to infinity since any number of interruptions is possible. The different interruption types can occur in any order, which is captured by the multinomial coefficient in the probability mass function (pmf) in (17).

2. Holds true for C^* due to the first part of Assumption 4, and because $\mathrm{EIFS} - \mathrm{DIFS} = \mathrm{SIFS} + t_{ack}$ (see [1]).

Next, we address the second term in (8), $A^{(k)}$. Since the number of backoff intervals that the tagged station experiences depends on the number of retransmissions, the value of $A^{(k)}$ strongly depends on the number of retransmissions. The number of retransmissions before success takes a truncated geometric distribution with pmf $\eta_k c_k^i$ for $i = 0, \ldots, R-1$. We can therefore write

$$A^{(k)} = A_i^{(k)} \quad \text{w.p.} \quad \eta_k c_k^i, \tag{18}$$

where $i = 0, \ldots, R-1$. The r.v. $A_i^{(k)}$ is comprised of i collisions involving the tagged station, $i+1$ backoff intervals and the interruptions to them. It can be expressed as

$$A_i^{(k)} = \sum_{j=0}^{i} B_{i,j}^{(k)} + \sum_{j=1}^{i} C_{i,j}^{(k)}, \tag{19}$$

where $B_{i,j}^{(k)}$ represents the backoff intervals and the interruptions, and $C_{i,j}^{(k)}$ represents the channel occupancy of a collision involving the tagged station. The r.v.'s $C_{i,j}^{(k)}$ are all i.i.d. and $B_{i,j}^{(k)}$ are i.i.d. in the index i.

For uniform, fixed length packets, we have

$$C^{(k)} = t_{data} + \mathrm{SIFS} + t_{ack} + \epsilon^{(k)}, \tag{20}$$

where the i, j subscripts are suppressed for notational clarity.

The scope of $B_{i,j}^{(k)}$ is defined by a backoff interval that takes a discrete uniform distribution. In EDCA, each slot of the backoff interval can be interrupted at most once with certain probabilities, either by a successful transmission from a nontagged station, or by a collision involving the nontagged stations. Each interruption causes the backoff timer to be frozen, and after the channel becomes idle again, the backoff process resumes from the next slot. Based on this, for any i, we can express B_j as a random sum

$$B_j^{(k)} = \sum_{n=1}^{U_j^{(k)}} Y_n^{(k)}, \tag{21}$$

where $Y_n^{(k)}$ is i.i.d. and represents the interruption to the nth backoff slot, and $U_j^{(k)}$ is the backoff interval given by (5).

In the following, we suppress the index n from $Y_n^{(k)}$ for clarity. If no other station transmits, $Y^{(k)}$ is equal to the duration of a slot time t_{slot}. If there is only one transmission, it is equal to the channel occupancy of a successful transmission, denoted as $G^{(k)}$. When more than one nontagged station attempts to transmit, $Y^{(k)}$ equals the channel occupancy of a collision involving nontagged stations, denoted by $H^{(k)}$. Hence, we obtain

$$Y^{(k)} = \begin{cases} t_{slot}, & w.p. \quad 1-c_k, \\ G^{(k)}, & w.p. \quad \gamma^{(k)}, \\ H^{(k)}, & w.p. \quad \nu^{(k)}, \end{cases} \tag{22}$$

where $\gamma^{(k)}$ and $\nu^{(k)}$ are the corresponding probabilities for successful transmissions and collisions, respectively. Like c_k, $\gamma^{(k)}$ and $\nu^{(k)}$ must be determined by averaging over slot classes: $\nu^{(k)} = c_k - \gamma^{(k)}$ and

$$\gamma^{(k)} = \sum_{j=k}^{J} \gamma^{(k)}(j) \frac{P(j)}{\sum_{i=k}^{J} P(i)},$$
$$\gamma^{(k)}(j) = (n_k - 1) p_k r_k^{n_k - 2} \prod_{\substack{i=1 \\ i \neq k}}^{j} r_i^{n_i} + r_k^{n_k - 1} \sum_{\substack{i=1 \\ i \neq k}}^{j} \left[n_i p_i r_i^{n_i - 1} \prod_{\substack{l=1 \\ l \neq k, l \neq i}}^{j} r_l^{n_l} \right]. \quad (23)$$

The first term in (23) is the probability that exactly one of the nontagged AC[k] stations transmits and no other station transmits; the second term is the sum of the probabilities that exactly one of the AC[i] ($i \neq k$) stations transmits and no other station transmits. For uniform, fixed length packets

$$G^{(k)} = H^{(k)} = t_{data} + \text{SIFS} + t_{ack} + \epsilon^{(k)}. \quad (24)$$

2.3 Generating Function

Now we derive the generating function of the distribution of the access delay for the case $\text{TXOP}_i = 0 (i = 1, \ldots, J)$, using the analysis of the previous section. We use the following notational convention for a generating function: if X is a nonnegative, integer-valued r.v., then the generating function of the pmf of X is

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

Since the r.v.'s introduced in Section 2.2 are not always integer-valued, we transform them to integer-valued r.v.'s by defining a lattice with fine spacing δ, such that the values of all r.v.'s are concentrated on the lattice points, and then scaling δ to 1. In the sequel, we abuse the notation slightly by reusing the r.v. names that appear in Section 2.2 to refer to their integer-valued equivalents.

We immediately obtain an expression for the generating function of the access delay $\widehat{D^{(k)}}(z)$ from (8):

$$\widehat{D^{(k)}}(z) = \widehat{A^{(k)}}(z) \widehat{T^{(k)}}(z) \widehat{\epsilon^{(k)}}(z). \quad (25)$$

In the following, we suppress the superscript (k) from the generating functions for notational clarity. For the case of fixed length packets, we have

$$\widehat{T}(z) = z^{t_{data}/\delta}. \quad (26)$$

Based on (18) and (19), we can write $\widehat{A}(z)$ as

$$\widehat{A}(z) = \sum_{i=0}^{R-1} \eta_k c_k^i \widehat{C}(z)^i \prod_{j=0}^{i} \widehat{B}_j(z). \quad (27)$$

It follows from (20) that

$$\widehat{C}(z) = \widehat{\epsilon}(z) z^{\omega}, \quad (28)$$

where ω is an integer constant defined by $\omega = (t_{data} + \text{SIFS} + t_{ack})/\delta$.

From (21), the generating function of $B_j^{(k)}$ is given by

$$\widehat{B}_j(z) = \widehat{U}_j\left(\widehat{Y}(z)\right). \quad (29)$$

Equation (5) yields

$$\widehat{U}_j(z) = \begin{cases} \frac{1 - z^{f(j)}}{f(j)(1-z)}, & \text{for} \quad j = 0, \ldots, m_k - 1, \\ \frac{1 - z^{f(m_k)}}{f(m_k)(1-z)}, & \text{for} \quad j = m_k, \ldots, R - 1, \end{cases}$$

where $f(j) = \langle \beta_k^j \mathrm{W}_k \rangle$. From (22), it follows that

$$\widehat{Y}(z) = (1 - c_k) z^{t_{slot}/\delta} + \gamma \widehat{G}(z) + \nu \widehat{H}(z), \quad (30)$$

where it is easy to obtain from (24) that

$$\widehat{G}(z) = \widehat{H}(z) = \widehat{\epsilon}(z) z^{\omega}. \quad (31)$$

The next step is to find $\widehat{\epsilon}(z)$. For the highest priority class, AC[1], we have that

$$\widehat{\epsilon}(z) = z^{\text{AIFS}_1/\delta}. \quad (32)$$

For other classes, $\widehat{\epsilon}(z)$ can be derived from (17) by invoking the multinomial theorem:

$$\widehat{\epsilon}(z) = \frac{z^{\text{AIFS}_k/\delta} s}{1 - \sum_{l=1}^{h} z^{t_l/\delta} \mu_l}. \quad (33)$$

For fixed length packets, we find that

$$t_l = \text{AIFS}_1 + (l - 1) t_{slot} + T^*.$$

Thus, the generating function of the pmf of the access delay can be derived from (25)-(33). In the numerical experiments reported in Section 4.1, we deal with the generating function of 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 $\widehat{D}(z)$ using

$$\widehat{D_c}(z) = \frac{1 - \widehat{D}(z)}{1 - z}. \quad (34)$$

The analytical distribution results reported in Section 4 are obtained by numerically inverting (34). We use the LATTICE-POISSON numerical inversion algorithm [23].

2.4 Mean and Standard Deviation

In this section, we derive the mean and standard deviation of the access delay for the case $\text{TXOP}_i = 0 (i = 1, \ldots, J)$. We denote the mean and the standard deviation by $\text{E}[D^{(k)}]$ and $\text{S}[D^{(k)}]$, respectively. Referring to (8), since $A^{(k)}$, $T^{(k)}$, and $\epsilon^{(k)}$ are independent, we can write

$$\text{E}\left[D^{(k)}\right] = \text{E}\left[\epsilon^{(k)}\right] + \text{E}\left[A^{(k)}\right] + \text{E}\left[T^{(k)}\right],$$
$$\text{S}\left[D^{(k)}\right] = \sqrt{\text{V}[\epsilon^{(k)}] + \text{V}[A^{(k)}] + \text{V}[T^{(k)}]},$$

where V[.] denotes the variance.

In the case of fixed length packets, we have

$$\text{E}\left[T^{(k)}\right] = t_{data}, \quad \text{V}\left[T^{(k)}\right] = 0.$$

For AC[1], it always holds that

$$\text{E}\left[\epsilon^{(1)}\right] = \text{AIFS}_1, \quad \text{V}\left[\epsilon^{(1)}\right] = 0.$$

For AC[k] ($k > 1$), the mean and variance of $\epsilon^{(k)}$ can be found from (17):

$$\mathrm{E}\left[\epsilon^{(k)}\right] = \mathrm{AIFS}_k + \frac{\sum_{l=1}^{h^{(k)}} \mu_l t_l}{1-\sum_{l=1}^{h^{(k)}} \mu_l},$$
$$\mathrm{V}\left[\epsilon^{(k)}\right] = \frac{\left(\sum_{l=1}^{h^{(k)}} \mu_l t_l\right)^2}{\left(1-\sum_{l=1}^{h^{(k)}} \mu_l\right)^2} + \frac{\sum_{l=1}^{h^{(k)}} \mu_l t_l^2}{1-\sum_{l=1}^{h^{(k)}} \mu_l}.$$

From (18), we can write $\mathrm{E}[A^{(k)}]$ and $\mathrm{V}[A^{(k)}]$ as

$$\mathrm{E}\left[A^{(k)}\right] = \sum_{i=0}^{R-1} \eta_k c_k^i \mathrm{E}\left[A_i^{(k)}\right],$$
$$\mathrm{V}\left[A^{(k)}\right] = \sum_{i=0}^{R-1} \eta_k c_k^i \left(\mathrm{V}\left[A_i^{(k)}\right] + \left(\mathrm{E}\left[A_i^{(k)}\right] - \mathrm{E}\left[A^{(k)}\right]\right)^2\right),$$

where, from (19), we have

$$\mathrm{E}\left[A_i^{(k)}\right] = \sum_{j=0}^{i} \mathrm{E}\left[B_j^{(k)}\right] + i\mathrm{E}\left[C^{(k)}\right],$$
$$\mathrm{V}\left[A_i^{(k)}\right] = \sum_{j=0}^{i} \mathrm{V}\left[B_j^{(k)}\right] + i\mathrm{V}\left[C^{(k)}\right].$$

For uniform, fixed length packets, it follows from (20) that

$$\mathrm{E}\left[C^{(k)}\right] = t_{data} + \mathrm{SIFS} + t_{ack} + \mathrm{E}\left[\epsilon^{(k)}\right],$$
$$\mathrm{V}\left[C^{(k)}\right] = \mathrm{V}\left[\epsilon^{(k)}\right].$$

The mean and variance of $B_j^{(k)}$ can be obtained from (21):

$$\mathrm{E}\left[B_j^{(k)}\right] = \mathrm{E}\left[U_j^{(k)}\right]\mathrm{E}\left[Y^{(k)}\right],$$
$$\mathrm{V}\left[B_j^{(k)}\right] = \mathrm{E}\left[U_j^{(k)}\right]\mathrm{V}\left[Y^{(k)}\right] + \mathrm{E}\left[Y^{(k)}\right]^2\mathrm{V}\left[U_j^{(k)}\right].$$

The mean of $U_j^{(k)}$ was given in (6). From (5), it is straightforward to show that

$$\mathrm{V}\left[U_j^{(k)}\right] = \begin{cases} \frac{1}{12}\left(\left\langle \beta_k^j \mathrm{W}_k \right\rangle^2 - 1\right), & \text{for } j = 0, \ldots, m_k - 1, \\ \frac{1}{12}\left(\left\langle \beta_k^{m_k} \mathrm{W}_k \right\rangle^2 - 1\right), & \text{for } j = m_k, \ldots, R-1. \end{cases} \quad (35)$$

It can be seen from (22) that the distribution of $Y^{(k)}$ is a simple mixture, so the mean and variance can be written as in (36). For the case of uniform, fixed length packets, we have

$$\mathrm{E}\left[G^{(k)}\right] = \mathrm{E}\left[H^{(k)}\right] = t_{data} + \mathrm{SIFS} + t_{ack} + \mathrm{E}\left[\epsilon^{(k)}\right],$$
$$\mathrm{V}\left[G^{(k)}\right] = \mathrm{V}\left[H^{(k)}\right] = \mathrm{V}\left[\epsilon^{(k)}\right].$$

Based on the equations above, the mean and variance of D can be obtained as in (37) and (38).

$$\mathrm{E}\left[Y^{(k)}\right] = (1-c_k)t_{slot} + \gamma^{(k)}\mathrm{E}\left[G^{(k)}\right] + \nu^{(k)}\mathrm{E}\left[H^{(k)}\right], \quad (36)$$

$$\mathrm{E}\left[D^{(k)}\right] = \eta_k \sum_{i=0}^{R-1} c_k^i \left\{ \mathrm{E}\left[Y^{(k)}\right] \sum_{j=0}^{i} \mathrm{E}\left[U_j^{(k)}\right] + i\mathrm{E}\left[C^{(k)}\right] \right\} + \mathrm{E}\left[T^{(k)}\right] + \mathrm{E}\left[\epsilon^{(k)}\right], \quad (37)$$

$$\mathrm{V}\left[D^{(k)}\right] = \eta_k \sum_{i=0}^{R-1} c_k^i \left\{ \sum_{j=0}^{i} \left(\mathrm{E}\left[U_j^{(k)}\right]\mathrm{V}\left[Y^{(k)}\right] + \mathrm{E}\left[Y^{(k)}\right]^2 \mathrm{V}\left[U_j^{(k)}\right] \right) + i\mathrm{V}\left[C^{(k)}\right] + \left(\mathrm{E}\left[Y^{(k)}\right] \sum_{j=0}^{i} \mathrm{E}\left[U_j^{(k)}\right] + i\mathrm{E}\left[C^{(k)}\right] - \mathrm{E}\left[A^{(k)}\right] \right)^2 \right\} + \mathrm{V}\left[T^{(k)}\right] + \mathrm{V}\left[\epsilon^{(k)}\right]. \quad (38)$$

2.5 TXOP Model

We analyze the access delay when differentiation by TXOP is configured. If $\mathrm{TXOP}_k > 0$ and an AC[k] station obtains the channel, it will be permitted to transmit a sequence of data packets in the time duration defined by TXOP_k. Since successive DATA-ACK exchanges are separated only by SIFS, collisions cannot occur except to the first transmitted packet.

Let us assume that the value of TXOP_k allows the sending of $N_k \geq 1$ consecutive packets. We denote the delay experienced by the N_k packets as $D_1^{(k)}, D_2^{(k)}, \ldots, D_{N_k}^{(k)}$, respectively. The MAC access delay for AC[k] can be expressed as

$$D^{(k)} = \begin{cases} D_1^{(k)}, & w.p. \quad 1/N_k, \\ D_2^{(k)}, & w.p. \quad 1/N_k, \\ \ldots & \\ D_{N_k}^{(k)}, & w.p. \quad 1/N_k, \end{cases} \quad (39)$$

where for $i = 2, 3, \ldots, N_k$, we have that

$$D_i^{(k)} = \mathrm{SIFS} + t_{data}, \quad (40)$$

and $D_1^{(k)}$ can be obtained in a similar way to that described in Section 2.2, using

$$D_1^{(k)} = \epsilon^{(k)} + A^{(k)} + t_{data}, \quad (41)$$

but with differences in some components of $\epsilon^{(k)}$ and $A^{(k)}$. The differences arise because the transmission durations are now extended and can vary between classes. Here, we demonstrate the constructions for them.

Clearly, $\epsilon^{(1)} = \mathrm{AIFS}_1$. An expression for $\epsilon^{(k)}(k > 1)$ can be obtained using (9)-(17), but with modifications to the expressions for X_i to separately account for different transmission durations between classes:

$$X_i = \begin{cases} T_l^*, & \text{w.p.} \quad \rho_l(i),\ 1 \leq l \leq \varphi(i), \\ C^*, & \text{w.p.} \quad 1 - \sum_{l=1}^{\varphi(i)} \rho_l(i), \end{cases}$$

where T_l^* is the channel occupancy of a successful transmission from an AC[l] station, and C^* is the channel occupancy of a collision involving any higher priority stations. The $\rho_l(i)$ is the probability of a successful

transmission. When all data packets in the system are of uniform, fixed length, we have

$$T_l^* = \Delta_l + \text{SIFS} + t_{ack},$$
$$C^* = t_{data} + \text{SIFS} + t_{ack}.$$

The term Δ_l is the successful transmission time of the N_l consecutive packets from an AC[l] station ($l \le \varphi(i)$), and is given by

$$\Delta_l = t_{data} + (N_l - 1)[2\text{SIFS} + t_{ack} + t_{data}].$$

The probabilities $\rho_l(i)$'s are obtained as

$$\rho_l(i) = \frac{n_l p_l r_l^{n_l-1} \prod_{\substack{j=1 \\ j\neq l}}^{\varphi(i)} r_j^{n_j}}{1 - \prod_{j=1}^{\varphi(i)} r_j^{n_j}}.$$

An expression for $A^{(k)}$ can be obtained using (18)-(21), together with the following modifications to $Y^{(k)}$:

$$Y^{(k)} = \begin{cases} t_{slot}, & w.p. \quad 1 - c_k, \\ G_l^{(k)}, & w.p. \quad \gamma_l^{(k)},\ l = 1, \ldots, J, \\ H^{(k)}, & w.p. \quad \nu^{(k)}, \end{cases}$$

where $G_l^{(k)}$ represents the channel occupancy of a successful transmission from an AC[l] station, and $H^{(k)}$ is the channel occupancy of a collision involving nontagged stations. In the case of uniform, fixed packet lengths, we have

$$G_l^{(k)} = \Delta_l + \text{SIFS} + t_{ack} + \epsilon^{(k)},$$
$$H^{(k)} = t_{data} + \text{SIFS} + t_{ack} + \epsilon^{(k)}.$$

The $\nu^{(k)}$ is obtained from $\nu^{(k)} = c_k - \sum_{l=1}^{J} \gamma_l^{(k)}$, and $\gamma_l^{(k)}$ from

$$\gamma_l^{(k)} = \sum_{j=\max(k,l)}^{J} \gamma_l^{(k)}(j) \frac{P(j)}{\sum_{i=k}^{J} P(i)}.$$

Here, the $\max$ function appears because the tagged AC[k] station can only decrement its backoff counter in slot class k or higher, and because AC[l] stations can only transmit in slot class l or higher. The probabilities $\gamma_l^{(k)}(j)$ are given by

$$\gamma_l^{(k)}(j) = \begin{cases} r_k^{n_k-1} n_l p_l r_l^{n_l-1} \displaystyle\prod_{i=1, i\neq k, i\neq l}^{j} r_i^{n_i}, & \text{for } l \neq k, \\ (n_k - 1) p_k r_k^{n_k-2} \displaystyle\prod_{i=1, i\neq k}^{j} r_i^{n_i}, & \text{for } l = k. \end{cases}$$

From expressions (40) and (41), the mean, standard deviation, and generating function of the pmf of $D_i^{(k)}$ can be derived. For $i = 1$, they are obtained in the same way as described in Section 2.4; for $i = 2, 3, \ldots, N_k$, it follows that

$$\mathrm{E}\left[D_i^{(k)}\right] = \text{SIFS} + t_{data},$$
$$\mathrm{V}\left[D_i^{(k)}\right] = 0,$$
$$\widehat{D_i^{(k)}}(z) = z^{(\text{SIFS}+t_{data})/\delta}.$$

Finally, the mean, standard deviation, and generating function of the pmf of $D^{(k)}$ follow from (39) as

$$\mathrm{E}\left[D^{(k)}\right] = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathrm{E}\left[D_i^{(k)}\right],$$
$$\mathrm{S}\left[D^{(k)}\right] = \sqrt{\frac{1}{N_k} \sum_{i=1}^{N_k} \left[\mathrm{V}\left[D_i^{(k)}\right] + \left(\mathrm{E}\left[D_i^{(k)}\right] - \mathrm{E}[D^{(k)}]\right)^2\right]}, \quad (42)$$
$$\widehat{D^{(k)}}(z) = \frac{1}{N_k} \sum_{i=1}^{N_k} \widehat{D_i^{(k)}}(z).$$

3 Asymptotic Analysis and Approximations

The expressions for the delay metrics found in Section 2 are accurate (as we demonstrate in Section 4.1) but their complexity obscures the influence of individual parameters and may also discourage their use. In this section, we strip away less essential details of the model to find simplified expressions for the mean and standard deviation that apply under various conditions. Using asymptotic analysis, we find the mean delay when $m = R = \infty$ under CWmin, AIFS, β, and TXOP differentiation. Then, to address the case of finite m and R, we develop approximations for both the mean and standard deviation. To facilitate the derivations of the asymptotics and approximations, we ignore the rounding operations that appear in (6) and (35), and we assume that data packets have a uniform, fixed length.

We consider a network with two classes of ACs, and refer to the high and low priority ACs as AC[1] and AC[2], respectively. Our aim is to find simplified expressions for $\mathrm{E}[D^{(k)}]$ and $\mathrm{V}[D^{(k)}]$, $k = 1, 2$. We also seek simple expressions for the *mean and standard deviation ratios*, which we define as $\theta_m := \mathrm{E}[D^{(2)}]/\mathrm{E}[D^{(1)}]$ and $\theta_s := \mathrm{S}[D^{(2)}]/\mathrm{S}[D^{(1)}]$, respectively. These moment ratios are useful metrics for quantifying the level of differentiation achieved.

3.1 Asymptotic Analysis

We study the asymptotic mean delay when $n \to \infty$. To obtain meaningful results, we assume $m = R = \infty$. The numbers of AC[1] and AC[2] stations are given by $n_1 = \alpha n$ and $n_2 = (1 - \alpha)n$, respectively, where $0 < \alpha < 1$. Ramaiyan et al. [20] previously studied asymptotic results for throughput ratios under the same conditions, and we make use of some of their intermediate results.

3.1.1 TXOP = 0

From the expression for the mean delay in (37), when $R = \infty$, we obtain

$$\mathrm{E}\left[D^{(k)}\right] = \frac{(1 - c_k)t_{slot} + c_k \mathrm{E}\left[C^{(k)}\right]}{p_k(1 - c_k)} + \frac{c_k \mathrm{E}\left[C^{(k)}\right]}{1 - c_k} + t_{data} + \mathrm{E}\left[\epsilon^{(k)}\right]. \quad (43)$$

The following lemmas and theorem summarize asymptotic results for differentiation by individual parameters:

$\mathrm{CW}_{\min}$ differentiation.

Lemma 1. *For $m = R = \infty$, when the service differentiation is provided by $\mathrm{CW}_{\min}$ with W_1, $W_2 \gg 1$, $\theta_m \to \frac{W_2 - 2\beta}{W_1 - 2\beta}$ as $n \to \infty$.*

Proof. It is shown in [20] that when $m = R = \infty$, for $k = 1, 2,$ we have

$$\lim_{n\to\infty} c_k \uparrow \frac{1}{\beta}, \quad \lim_{n\to\infty} p_k \downarrow 0. \tag{44}$$

It can also be shown that when $W_1, W_2 \gg 1$

$$p_k = \frac{1-\beta c_k}{\frac{W_k}{2}(1-c_k)}, \quad 0 \le c_k < \frac{1}{\beta}. \tag{45}$$

Taking the limit of θ_m using (43) and applying (44) and (45) leads to the result. □

AIFS differentiation.

Lemma 2. *For $m = R = \infty$, when the service differentiation is provided by AIFS*

$$\lim_{n\to\infty} \mathrm{E}\left[D^{(1)}\right] = \frac{n_1\left[(\beta-1)t_{slot} + \mathrm{E}\left[C^{(1)}\right]\right]}{(\beta-1)\ln\frac{\beta}{\beta-1}},$$

and $\theta_m \to \infty$ as $n \to \infty$.

Proof. In [20], it is shown that for AIFS differentiation, when $m = R = \infty$, (44) still holds, and, in addition

$$\begin{aligned} &\lim_{n\to\infty} n_1 p_1 \uparrow \ln\frac{\beta}{\beta-1}, \\ &\lim_{n\to\infty} n_2 p_2 = 0. \end{aligned} \tag{46}$$

Taking the limit of $\mathrm{E}[D^{(1)}]$ using (43) and applying (44) and (46) yields the asymptotic result for $\mathrm{E}[D^{(1)}]$. Similarly, it can be shown that $\mathrm{E}[D^{(2)}] \to \infty$ as $n \to \infty$, which leads to the result for θ_m. □

β differentiation.

Lemma 3. *For $m = R = \infty$, when the service differentiation is provided by β*

$$\lim_{n\to\infty} \mathrm{E}\left[D^{(1)}\right] = \frac{n_1[(\beta_1-1)t_{slot} + \mathrm{E}[C]]}{(\beta_1-1)\ln\frac{\beta_1}{\beta_1-1}},$$

and $\theta_m \to \infty$ as $n \to \infty$.

Proof. The proof follows similar lines to that of Lemma 2, using the following results from [20]:

$$\begin{aligned} &\lim_{n\to\infty} c_1 \uparrow \frac{1}{\beta_1} \quad \text{and} \quad \lim_{n\to\infty} c_2 \uparrow \frac{1}{\beta_1}, \\ &\lim_{n\to\infty} n_1 p_1 \uparrow \ln\frac{\beta_1}{\beta_1-1}, \\ &\lim_{n\to\infty} n_2 p_2 = 0. \end{aligned}$$

□

The asymptotic θ_m under $\mathrm{CW}_{\min}$ differentiation approaches the ratio of the AC initial windows if the initial windows are large ($W_k \gg 2\beta, k = 1, 2$). For AIFS and β differentiation, we only find the trivial asymptotic limit for θ_m, since the asymptotic mean delay of AC[1] stations is a linear function of n, but the asymptotic mean delay of AC[2] stations grows faster than linearly with n. Ramaiyan et al. [20] obtained similar asymptotic results for the throughput ratio, but did not provide the asymptotic results for the high priority class.

3.1.2 TXOP *Differentiation*

Theorem 1. *For $m = R = \infty$, $\theta_m \to \frac{N_1}{N_2}$ as $n \to \infty$.*

Proof. As all parameters except TXOP limit are identical for the two classes, we have

$$\begin{aligned} &p_1 = p_2 = p \quad \text{and} \quad \lim_{n\to\infty} p \downarrow 0, \\ &c_1 = c_2 = c \quad \text{and} \quad \lim_{n\to\infty} c \uparrow \frac{1}{\beta}, \\ &\lim_{n\to\infty} np \quad \uparrow \quad \ln\frac{\beta}{\beta-1}. \end{aligned} \tag{47}$$

From (42), (43), and (47), we have

$$\lim_{n\to\infty} \mathrm{E}\left[D^{(k)}\right] = \frac{n((\beta-1)t_{slot} + \mathrm{E}[C])}{N_k(\beta-1)\ln\frac{\beta}{\beta-1}}.$$

Taking the ratio of the asymptotic mean delays leads to the result. □

The asymptotic θ_m under TXOP differentiation is very simple and depends only on the value of the TXOP limit parameters. This result has not been observed previously in the literature.

3.2 Approximations

The approximations are derived under the assumption of finite $m = R$. To facilitate simplification, we make the following additional assumptions:

1. $n = n_1 + n_2$ is large (high load), so that c_1, c_2 approach 1 and p_1, p_2 approach 0.
2. W_1, $W_2 \gg 1$.
3. $t_{data} \gg t_{slot}$ and $t_{data} \gg h t_{slot}$.
4. R and β_1, β_2 are sufficiently large.

Assumptions 2 and 3 will hold for typical settings of these parameters. Regarding assumption 4, our numerical experience is that for $R = 7$, $\beta \ge 2$ is large enough to make the approximation suitably accurate (see Section 4.2). For simplicity, we drop the class index k from the notation in the following when there is no risk of ambiguity.

3.2.1 TXOP = 0

We consider differentiation by up to three parameters, namely CWmin, AIFS, and β. Under the assumptions listed previously, we obtain the following approximations:

$$\mathrm{E}[D] \approx \frac{c\beta\Gamma}{p(\beta-1)q^h}, \tag{48}$$

$$\mathrm{V}[D] \approx \frac{c^2W^2\Gamma^2}{q^{2h}} \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \tag{49}$$

where $\Gamma = \mathrm{AIFS}_1 + t_{data} + \mathrm{SIFS} + t_{ack}$, $q = (1-p_1)^{n_1}$ and $h = 0$ for class 1 and $h = (\mathrm{AIFS}_2 - \mathrm{AIFS}_1)/t_{slot}$ for class 2. The derivations of (48) and (49) are given in the Appendix.

Straightforwardly, the moment ratios are given by

$$\begin{aligned} \theta_m &\approx \frac{p_1 c_2}{p_2 c_1 q^{h^{(2)}}} \frac{\beta_2(\beta_1-1)}{\beta_1(\beta_2-1)} \\ &\approx \frac{\beta_2(\beta_1-1)}{\beta_1(\beta_2-1)} \frac{W_2 c_2 \sum_{i=0}^{R-1} \eta_2(\beta_2 c_2)^i}{q^{h^{(2)}} W_1 c_1 \sum_{i=0}^{R-1} \eta_1(\beta_1 c_1)^i}, \end{aligned} \tag{50}$$

TABLE 1
MAC and PHYS Parameters for 802.11b System

Parameter	Symbol	Value
SIFS	SIFS	10 μs
Slot time	t_{slot}	20 μs
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
Data rate	r_{data}	11 Mbps
Control rate	r_{ctrl}	1 Mbps

$$\theta_s \approx \frac{c_2 W_2 \beta_2 (\beta_1 - 1)}{c_1 W_1 \beta_1 (\beta_2 - 1) q^{h^{(2)}}} \times \sqrt{\frac{(2\beta_2 + 1)(\beta_1 + 1)}{(2\beta_1 + 1)(\beta_2 + 1)} \frac{\sum_{i=0}^{R-1} \eta_2 c_2^i \beta_2^{2i}}{\sum_{i=0}^{R-1} \eta_1 c_1^i \beta_1^{2i}}}. \quad (51)$$

For CWmin only differentiation, $\beta_1 = \beta_2$, $h^{(2)} = 0$, and $c_1 \approx c_2$ for sufficiently large n_1 and n_2 [5], so both (50) and (51) simplify to W_2/W_1.

3.2.2 TXOP *Differentiation*

We present approximations for TXOP only differentiation. From Section 2.1, we observe that for the TXOP differentiation only case, $c_1 = c_2 = c$ and $p_1 = p_2 = p$.

We obtain the following approximations:

$$\mathrm{E}[D] \approx \frac{[c + (n_1 N_1 + n_2 N_2 - n) p (1-p)^n] \beta \Gamma}{N p (\beta - 1)}, \quad (52)$$

$$\mathrm{V}[D] \approx \frac{[c + (n_1 N_1 + n_2 N_2 - n) p (1-p)^n]^2 W^2 \Gamma^2}{N} \times \frac{(2\beta + 1)\beta^2}{6(\beta + 1)(\beta - 1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \quad (53)$$

where the derivations appear in the Appendix. It follows that the approximate moment ratios are then given by

$$\theta_m \approx \frac{N_1}{N_2}, \quad \theta_s \approx \sqrt{\frac{N_1}{N_2}}. \quad (54)$$

4 Results and Discussion

This section has three objectives: 1) to compare numerical results obtained from the analysis of Section 2 with simulation in order to confirm the accuracy of the model, 2) to utilize the model to study the effectiveness of the various differentiation mechanisms for service separation, and 3) to test the accuracy of the approximations presented in Section 3.2. The simulations were conducted using the ns-2 (version 2.28) simulator [24], combined with an EDCA module developed by TU-Berlin [25]. A detailed examination of the simulation code revealed some inconsistencies between the code and the IEEE 802.11e standard [2], and these were fixed. The main discrepancies were

- after the backoff counter is frozen, the remaining backoff time is incorrectly calculated, and
- a post-backoff is not initiated when a packet is discarded due to the retry limit being reached.

The simulated network topology comprised of n saturated stations sending data packets to an access point (AP) under ideal channel conditions (i.e., no transmission errors due to the wireless channel). User datagram protocol (UDP) packets were used with a fixed size of 1,000 bytes. The MAC and physical layer parameters were configured in accordance with the default values in IEEE 802.11b, as shown in Table 1.

TABLE 2
EDCA Parameters Used in Simulation

Scen.	$n_1 : n_2$	$\mathrm{W_k}$		$\mathrm{M_k}$		$\mathrm{AIFS_k}$ (μs)		$\mathrm{TXOP_k}$ (ms)	
1	1 : 2	16	32	1024	1024	50	50	0	0
2	1 : 2	32	32	1024	1024	50	70	0	0
3	1 : 1	8	16	512	1024	50	70	2.9	0

Accordingly, the durations for data and acknowledgment packet transmissions used in our analytical model are

$$t_{data} = t_{phys} + \frac{l_{mac} + l_{udpip} + l_{pay}}{r_{data}},$$
$$t_{ack} = t_{phys} + \frac{l_{ack}}{r_{ctrl}},$$

where l_{pay} is the UDP packet payload in bits. Propagation delays were ignored in the analytical model as they are several orders of magnitude smaller than the transmission times.

4.1 Validation

To corroborate the accuracy of the analysis of Section 2, we compare numerical values obtained from our model for the mean, standard deviation, and CCDF of the access delay with results obtained from simulation. For the analytical computation of the CCDF, we used a small lattice spacing $\delta = 10\ \mu$s to make the discretization error negligible, and used inversion parameters to give an inversion error no greater than 10^{-8}. The simulation results for the mean and standard deviation are plotted with 95 percent confidence intervals derived from five runs for each point in the graphs. In accordance with the standard [2], all numerical examples in this section use $R = 7$ and $\beta = 2$.

We start by considering two groups of stations, each with traffic belonging to a single AC, and we denote the number of stations of the high and low priority ACs by n_1 and n_2, respectively. Table 2 lists the $n_1 : n_2$ ratios and the differentiation parameters of three scenarios that were investigated. The first two scenarios test the differentiation achieved through only one parameter at a time, namely CWmin and AIFS, respectively.

The results (mean, standard deviation, and CCDF) for scenarios 1 and 2 are shown together in Fig. 2. The mean and standard deviation results in Figs. 2a and 2b are plotted against different total numbers of stations $n_1 + n_2$, while the CCDF results in Fig. 2c necessarily pertain to a specific n_1 and n_2. Observe that the analytical values are an excellent match with the simulation results. For the CCDF values, accuracy is maintained down to small tail probabilities.

In Fig. 3, we present results for the mean, standard deviation, and CCDF for the last scenario of Table 2. These results demonstrate that our model accurately predicts performance when all four differentiation mechanisms in the standard are activated, namely CWmin, CWmax, AIFS,

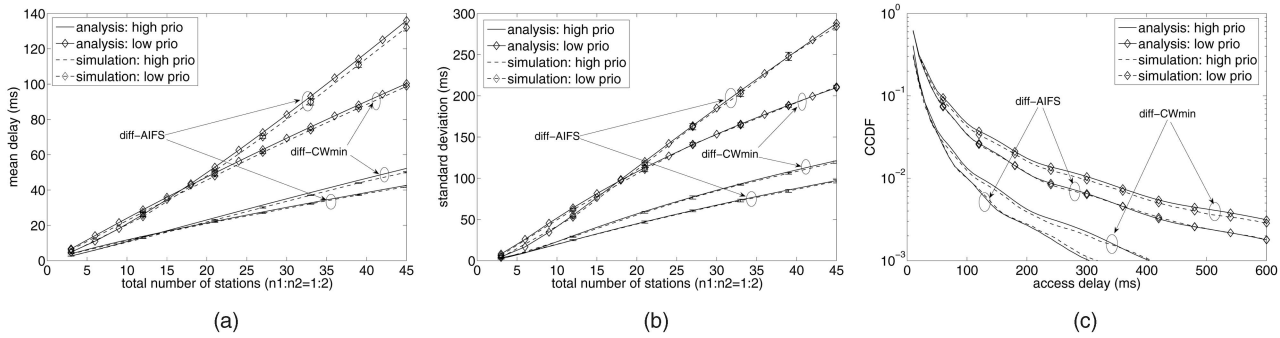

Fig. 2. Differentiation by CWmin or AIFS. (a) Mean access delay. (b) Standard deviation of access delay. (c) CCDF of access delay, $n_1 = 4$ and $n_2 = 8$.

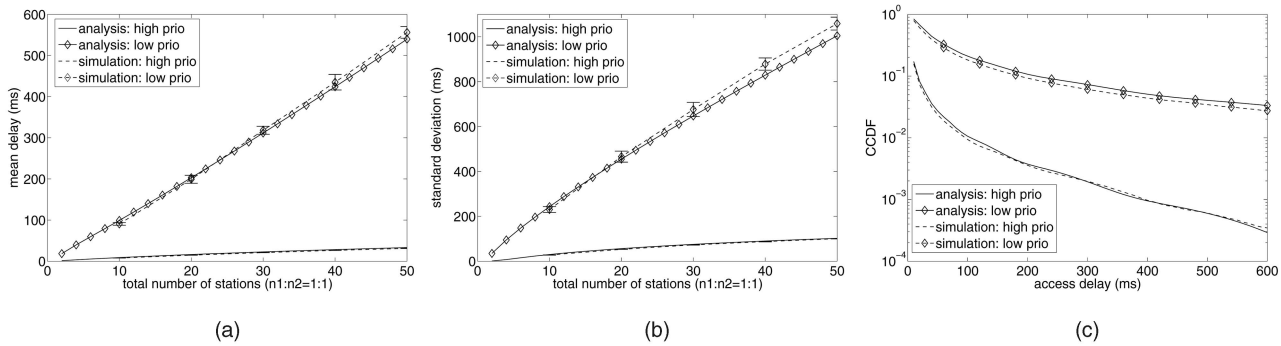

Fig. 3. Differentiation by CWmin, CWmax, AIFS, and TXOP limit. (a) Mean access delay. (b) Standard deviation of access delay. (c) CCDF of access delay, $n_1 = n_2 = 5$.

and TXOP limit. As one would expect, combining differentiation mechanisms leads to a greater degree of service separation between classes than using each mechanism individually.

To show that our analytical model is not restricted to just two ACs, we present results in Fig. 4 for an example with four ACs. The following parameters settings were used: $\mathrm{W}_k = \{8/8/32/32\}$ and $\mathrm{AIFS}_k = \{50/70/70/90\}$ μs. The values of other parameters were common for all classes: $\mathrm{M} = 1{,}024$ and $\mathrm{TXOP} = 0$.

In terms of a method for obtaining values of the distribution, the generating function analysis of Engelstad and Østerbø [13] comes closest in spirit to our approach. In Fig. 5, we plot the CCDFs obtained by numerically inverting our generating function and inverting the generating function derived in [13] for the saturation condition. The parameters were the same as in scenario 2 of Table 2, except that the AIFS of the lower priority stations was set to 90 μs. Our CCDF is a much better match with the simulations compared to the Engelstad CCDF, especially for the lower priority AC. The inaccuracy of the model in [13] stems from the fact that the authors use a coarse approximation technique to account for AIFS differentiation based on a simple scaling of the probability of detecting an idle slot [12], and do not include the additional delay caused by multiple interruptions to the AIFS of low priority stations.

Additional validation results omitted due to space limitations can be found in [21]. Our numerical experience is that the model maintains accuracy over a wide range of parameter values. However, for $\mathrm{CWmin} \leq 4$, the accuracy sometimes degrades. We attribute this to the multistability phenomenon described in [20], which gives rise to multiple solutions for the fixed-point.

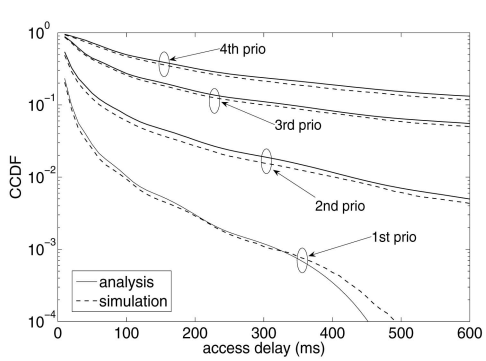

Fig. 4. CCDF of access delay: four classes, $n_1 = n_2 = n_3 = n_4 = 4$.

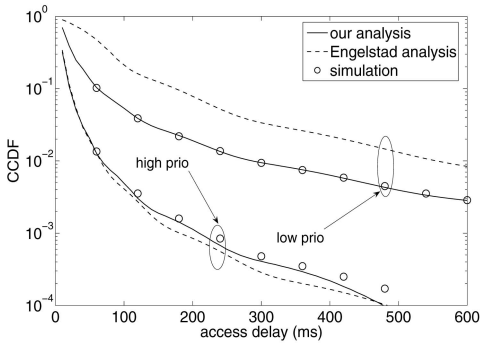

Fig. 5. CCDF of access delay: comparison with Engelstad result [13], $n_1 = 4$ and $n_2 = 8$.

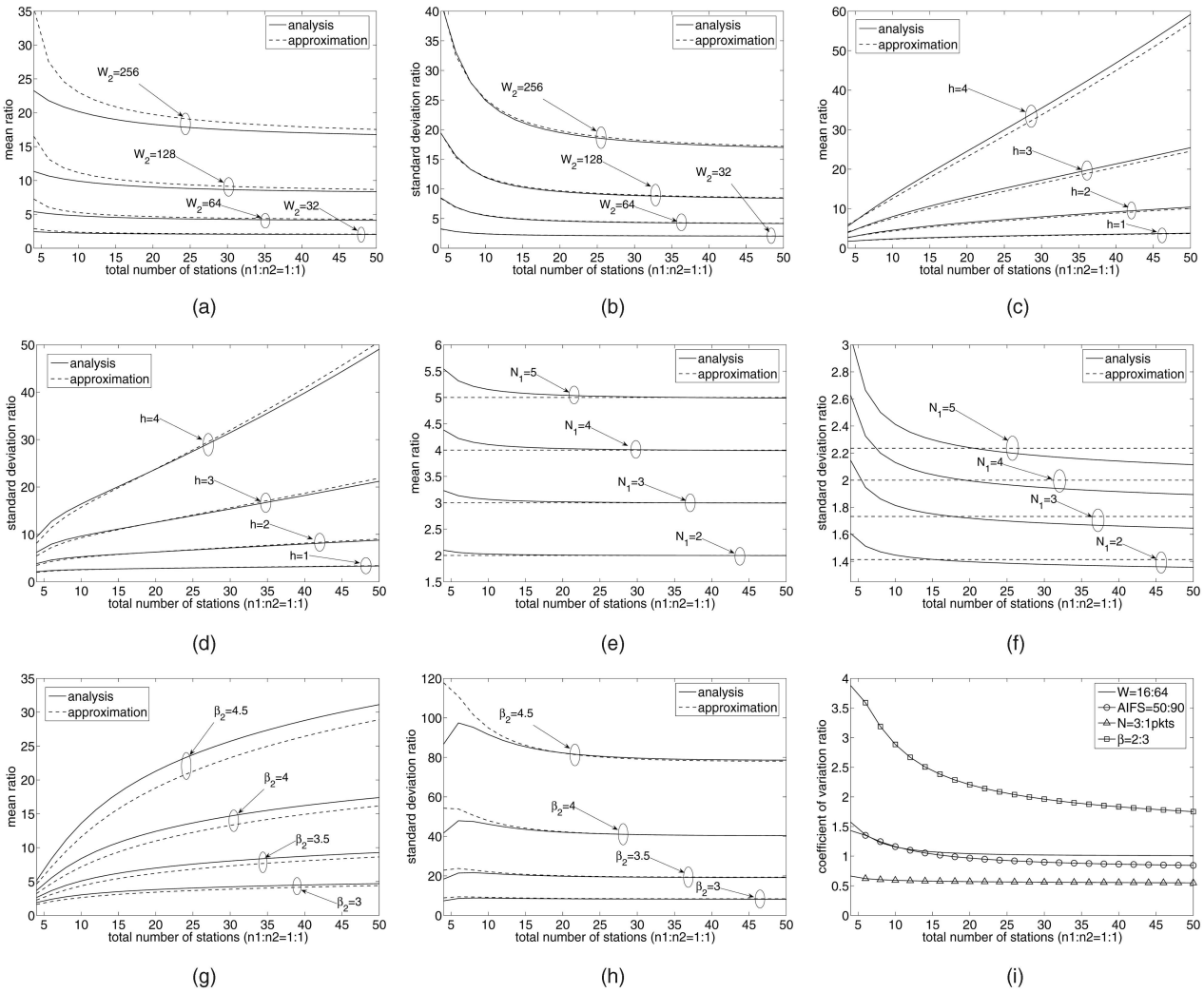

Fig. 6. Ratio metrics for different differentiation mechanisms. (a) Mean ratio for $\mathrm{CW}_{\min}$ differentiation. (b) Standard deviation ratio for $\mathrm{CW}_{\min}$ differentiation. (c) Mean ratio for AIFS differentiation. (d) Standard deviation ratio for AIFS differentiation. (e) Mean ratio for TXOP differentiation. (f) Standard deviation ratio for TXOP differentiation. (g) Mean ratio for β differentiation. (h) Standard deviation ratio for β differentiation. (i) Coefficient of variation ratio.

4.2 Comparison of Differentiation Mechanisms

Having established the validity of our analytical model, we now use it to quantify and compare the influence of each differentiation mechanism in greater detail. Concurrently, we investigate the accuracy of the approximations in Section 3.2. Since the approximations are derived under the assumption $m = R$, we fix $m = R = 7$ for all classes in the numerical examples in this section. We focus on service differentiation through AIFS, CWmin, TXOP limit, and β. We do not study CWmax differentiation explicitly, since a consequence of a fixed m is that any adjustment in CWmin or β leads to a corresponding adjustment in CWmax and vice versa. Therefore, CWmax differentiation occurs as byproduct of CWmin differentiation and β differentiation. These joint differentiation cases will be referred to as simply CWmin or β differentiation, since the relatively large value of m relegates CWmax to secondary importance.

Consider a setting with two ACs with equal numbers of stations, and define the following reference set of parameter values: $\{\mathrm{W}, \mathrm{AIFS}, \mathrm{TXOP}, \beta\} = \{16, 50\ \mu\mathrm{s}, 0, 2\}$. In the examples shown in this section, we impart service differentiation through one or more parameters by varying the relevant parameters of one AC away from the above reference settings, while maintaining all other parameters for both ACs at the reference settings. In each example, we will refer to the high and low priority ACs as $\mathrm{AC}[1]$ and $\mathrm{AC}[2]$, respectively. To measure the degree of service differentiation, we plot the moment ratios θ_m and θ_s. The approximations for θ_m and θ_s are computed using (54) for TXOP differentiation, and (50) and (51) for the other mechanisms.

The analytical and approximate moment ratios are presented in Fig. 6. The moment ratios under CWmin differentiation are illustrated in Figs. 6a and 6b. In this example, $\mathrm{W}_1 = 16$, while $\mathrm{W}_2 = 32, 64, 128, 256$. We see that θ_m and θ_s initially decrease before becoming largely insensitive to load. At high load, both ratios are roughly equal to the ratio of the two CWmin values, which is consistent with the asymptotic result in Section 3.1 and the observations made in Section 3.2. A consequence of a nonincreasing moment ratio is that high priority traffic may not be adequately protected under congestion. On the other hand, a constant ratio delivers predictability, which simplifies network planning and design.

Figs. 6c and 6d depict moment ratios when increasing levels of AIFS differentiation are applied. Specifically, AIFS_1 is maintained at 50 μs, while $\mathrm{AIFS}_2 = 70, 90, 110, 130$ μs. Observe that both the delay and standard deviation ratios grow as the total number of nodes in the network increases. That is, AIFS differentiation gives protection to high priority traffic by penalizing lower priority traffic when the contention level in the network increases. While this is essentially desirable, a negative ramification of this type of service separation is that it could lead to starvation for lower priority traffic under high load.

Results for TXOP differentiation are presented in Figs. 6e and 6f, where TXOP_2 is fixed at 0 and TXOP_1 is varied to permit the transmission of two, three, four, or five packets. The shapes of the TXOP curves are similar to those for CWmin differentiation, but TXOP yields greater predictability and finer-grained control of the level of differentiation.

In [3], it is stated that β differentiation was abandoned during the standardization process because its performance is similar to CWmin differentiation, though less effective. Figs. 6g and 6h show the moment ratios for β differentiation, where β_1 is fixed at 2 and $\beta_2 = 3, 3.5, 4, 4.5$. Comparison with Figs. 6a and 6b reveals that, contrary to the claims in [3], β differentiation is effective. It also yields dissimilar performance to CWmin differentiation; the mean ratio curves for β differentiation increase with load, while the standard deviation ratio curves are flatter than those for CWmin differentiation for small numbers of stations.

Figs. 6a, 6b, 6c, 6d, 6e, 6f, 6g, and 6h reveal that the approximations are accurate enough to capture the key trends in the service differentiation, except for the low load regime in some examples. In certain cases, such as the standard deviation ratios for CWmin, AIFS, and β differentiation, the agreement is excellent. The simplicity of the approximations compared to the complete analytical expressions makes them an attractive alternative for system design and configuration.

A further way to compare the differentiation mechanisms is to look at the *coefficients of variation* v_1 and v_2 of the delay distributions of AC[1] and AC[2]. The coefficient of variation of a probability distribution is the ratio of the standard deviation to the mean and is a measure of the dispersion relative to the mean. Ideally, we would like to have $v_1 < v_2$; that is, the delay of the high priority class should exhibit less dispersion than that of the low priority class. In Fig. 6i, we plot analytical curves of the coefficient of variation ratio v_2/v_1 for examples of each type of differentiation selected from the previous figures (note that $v_2/v_1 = \theta_s/\theta_m$). We can see that for $\mathrm{CW}_{\min}$ and AIFS differentiation, v_2/v_1 is approximately 1, while for TXOP differentiation, v_2/v_1 is always less than 1. In contrast, v_2/v_1 for β differentiation is greater than 1, so in this respect, β differentiation is superior to the other mechanisms.

5 Conclusion

In this paper, we have developed an accurate and versatile model for the MAC access delay in an IEEE 802.11e EDCA network under saturation. Using the model, we derive asymptotics and approximations for the mean and standard deviation of the access delay. The resultant expressions yield insights into the relative importance of different model parameters. Further simplification is achieved by forming the mean and standard deviation ratios; in particular, for CWmin differentiation, both the mean and standard deviation ratios can be approximated by the ratio of the minimum contention windows, while for TXOP differentiation, the mean and standard deviation ratios can be approximated by the inverse of the ratio of the burst sizes and its square root, respectively. We also use the model to study the effectiveness of CWmin, AIFS, TXOP, and β differentiation. We find that the AIFS mechanism gives protection to higher priority traffic under congestion. On the other hand, the CWmin and TXOP mechanisms give differentiation that is largely insensitive to the load which leads to fairly predictable behavior. Differentiation based on β has the desirable property of preserving priority in the delay dispersion relative to the mean.

Appendix A

In the following, to simplify the notation, we suppress the class index k when there is no risk of ambiguity.

A.1 Derivation of (48)

We approximate the mean delay as follows:

$$\mathrm{E}[D] \approx \mathrm{E}[A] \tag{55}$$

$$\approx \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]\mathrm{E}[Y], \tag{56}$$

where (55) follows because, when c is large, backoff windows become large and more interruptions occur, and so $\mathrm{E}[A]$ dominates the other terms. Similarly, (56) follows because $\sum_{j=0}^{i} \mathrm{E}[B_j] \gg i\mathrm{E}[C]$ when c is large and $\mathrm{E}[B_j] = \mathrm{E}[U_j]\mathrm{E}[Y]$. To simplify (56) further, we derive approximations for $\mathrm{E}[Y]$ and $\sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]$. We write $\mathrm{E}[Y]$ as

$$\mathrm{E}[Y] = (1-c)t_{slot} + c\mathrm{E}[G] \approx c\mathrm{E}[G] \approx \frac{c\Gamma}{q^h}, \tag{57}$$

where $\Gamma := \mathrm{AIFS}_1 + t_{data} + \mathrm{SIFS} + t_{ack}$, and $h := h^{(k)} = (\mathrm{AIFS}_k - \mathrm{AIFS}_1)/t_{slot}$. Equation (57) follows from $\mathrm{E}[G] \gg t_{slot}$ and $q^h \ll 1$ and the assumption that $ht_{slot} \ll t_{data}$.

We also have

$$\sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j] \approx \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \frac{W}{2}\beta^j \tag{58}$$

$$\approx \frac{1}{\beta - 1}\frac{W}{2}\beta \sum_{i=0}^{R-1} \eta c^i \beta^i, \tag{59}$$

where (58) follows under the assumption $W \gg 1$, (59) is obtained because for sufficiently large c and β, it can be shown that

$$\beta \sum_{i=0}^{R-1} \eta c^i \beta^i \gg 1. \tag{60}$$

Finally, by substituting (57) and (59) into (56), and using

$$\frac{1}{p} = \sum_{i=0}^{R-1} \eta c^i \mathrm{E}[U_i] \approx \frac{W}{2} \sum_{i=0}^{R-1} \eta c^i \beta^i, \tag{61}$$

we obtain (48).

A.2 Derivation of (49)

We approximate the variance as

$$\mathrm{V}[D] \approx \mathrm{V}[A] \tag{62}$$

$$\approx \sum_{i=0}^{R-1} \eta c^i \left\{ \sum_{j=0}^{i} \mathrm{V}[B_j] + \mathrm{E}[A_i]^2 \right\} - \mathrm{E}[A]^2 \tag{63}$$

$$\approx \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]\mathrm{V}[Y] + \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{V}[U_j]\mathrm{E}[Y]^2 + \sum_{i=0}^{R-1} \eta c^i \left(\sum_{j=0}^{i} \mathrm{E}[U_j]\mathrm{E}[Y] \right)^2 - \mathrm{E}[A]^2, \tag{64}$$

where (62) follows because $\mathrm{V}[A] \gg \mathrm{V}[\epsilon]$ (since c is large), (63) from $\sum_{j=0}^{i} \mathrm{V}[B_j] \gg \mathrm{V}[\epsilon]$.

We then approximate $\mathrm{V}[Y]$ and $\mathrm{V}[U_j]$ as

$$\mathrm{V}[Y] \approx c(1-c)\mathrm{E}[G]^2 + c\mathrm{V}[\epsilon] \tag{65}$$

$$\approx c(2-c)\mathrm{E}[G]^2, \tag{66}$$

$$\mathrm{V}[U_j] = \frac{\beta^{2j}W^2 - 1}{12} \approx \frac{W^2}{12}\beta^{2j}, \tag{67}$$

where, (65) is again because $\mathrm{E}[G] \gg t_{slot}$, (66) follows from $q^h \ll 1$, and (67) is due to $W \gg 1$. To further simplify (64), we note that for large R and sufficiently large β, it can be shown that

$$\sum_{i=0}^{R-1} \eta c^i \beta^{2i} \gg \sum_{i=0}^{R-1} \eta c^i \beta^i, \tag{68}$$

$$\sum_{i=0}^{R-1} \eta c^i \beta^{2i} \gg \left(\sum_{i=0}^{R-1} \eta c^i \beta^i \right)^2. \tag{69}$$

Based on (60), (68), and (69), and under the assumption that $W \gg 1$, we obtain

$$\sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{V}[U_j] \approx \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \frac{\beta^{2j}W^2}{12} \approx \frac{W^2\beta^2}{12(\beta^2-1)} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \tag{70}$$

$$\sum_{i=0}^{R-1} \eta c^i \left(\sum_{j=0}^{i} \mathrm{E}[U_j] \right)^2 \approx \sum_{i=0}^{R-1} \eta c^i \left(\sum_{j=0}^{i} \frac{W}{2}\beta^j \right)^2 \approx \frac{W^2\beta^2}{4(\beta-1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}. \tag{71}$$

Finally, substituting (56), (57), (59), (66), (67), (70), and (71) into (64), and using (68) and (69), we obtain

$$\mathrm{V}[D] \approx \frac{W^2\beta^2}{12(\beta^2-1)} c^2 \mathrm{E}[G]^2 \sum_{i=0}^{R-1} \eta c^i \beta^{2i} + \frac{W^2\beta^2}{4(\beta-1)^2} c^2 \mathrm{E}[G]^2 \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \tag{72}$$

$$\approx \frac{c^2 W^2 \Gamma^2}{q^{2h}} \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}. \tag{73}$$

A.3 Derivation of (52) and (53)

The mean delay is given by

$$\mathrm{E}[D] \approx \frac{\mathrm{E}[D_1]}{N} \tag{74}$$

$$\approx \frac{\mathrm{E}[Y] \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]}{N}, \tag{75}$$

where (74) follows because $\mathrm{E}[D_1] \gg (N-1)(\mathrm{SIFS} + t_{data})$ when c is sufficiently large, and (75) comes from (56). We approximate $\mathrm{E}[Y]$ as follows:

$$\mathrm{E}[Y] = (1-c)t_{slot} + \gamma_1 \mathrm{E}[G_1] + \gamma_2 \mathrm{E}[G_2] + \nu \mathrm{E}[H] \approx [c + (n_1 N_1 + n_2 N_2 - n)p(1-p)^n]\Gamma. \tag{76}$$

Substituting (59) and (76) into (75), and making use of (61), yields (52).

The variance of the access delay can be simplified as

$$\mathrm{V}[D] \approx \frac{\mathrm{V}[D_1] + \frac{N-1}{N}\mathrm{E}[D_1]^2}{N} \tag{77}$$

$$\approx \frac{\mathrm{V}[D_1] + \mathrm{E}[D_1]^2}{N} \tag{78}$$

$$\approx \frac{\sum_{i=0}^{R-1} \eta c^i \left\{ \sum_{j=0}^{i} \mathrm{V}[B_j] + \mathrm{E}[A_i]^2 \right\}}{N} \tag{79}$$

$$\approx \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2} \frac{W^2 \mathrm{E}[Y]^2}{N} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \tag{80}$$

where (77) is obtained using the fact that $\mathrm{E}[D_1] \gg t_{data} + \mathrm{SIFS}$, (78) follows from $N(\mathrm{V}[D_1] + \mathrm{E}[D_1]^2) \gg \mathrm{E}[D_1]^2$ for sufficiently large β, (79) follows from (55) and (63), and (80) is obtained via similar arguments that led from (63) to (73). Finally, substituting (76) into (80) leads to (53).

REFERENCES

[1] *IEEE 802.11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,* IEEE, 1999.

[2] *IEEE 802.11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements,* IEEE, 2005.

[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] T. Sakurai and H.L. Vu, "Access Delay of the IEEE 802.11 MAC Protocol under Saturation," *IEEE Trans. Wireless Comm.,* vol. 6, no. 5, pp. 1702-1710, 2007.

[5] D. Xu, T. Sakurai, and H.L. Vu, "An Analytical Model of MAC Access Delay in IEEE 802.11e EDCA," *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '06),* vol. 4, pp. 1938-1943, 2006.

[6] D. Xu, T. Sakurai, and H.L. Vu, "MAC Access Delay in IEEE 802.11e EDCA," *Proc. IEEE 64th Vehicular Technology Conf. (VTC '06),* Fall 2006.

[7] Y. Xiao, "Performance Analysis for Priority Schemes for IEEE 802.11 and IEEE 802.11e Wireless LANs," *IEEE Trans. Wireless Comm.,* vol. 4, no. 4, pp. 1506-1515, 2005.

[8] Z. Kong, D.H. Tsang, B. Bensaou, and D. Gao, "Performance Analysis of IEEE 802.11e Contention-Based Channel Access," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 10, pp. 2095-2106, 2004.

[9] X. Tao and S. Panwar, "Throughput and Delay Analysis for the IEEE 802.11e Enhanced Distributed Channel Access," *IEEE Trans. Comm.,* vol. 54, no. 4, pp. 596-603, 2006.

[10] T.-C. Tsai and M.-J. Wu, "An Analytical Model for IEEE 802.11e EDCA," *Proc. IEEE Int'l Conf. Comm. (ICC '05),* vol. 5, pp. 3474-3478, 2005.

[11] H. Zhu and I. Chlamtac, "Performance Analysis for IEEE 802.11e EDCF Service Differentiation," *IEEE Trans. Wireless Comm.,* vol. 4, no. 4, pp. 1779-1788, 2005.
[12] P.E. Engelstad and O.N. Østerbø, "Non-Saturation and Saturation Analysis of IEEE 802.11e EDCA with Starvation Prediction," *Proc. Eighth ACM Int'l Symp. Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWiM '05),* pp. 224-233, 2005.
[13] P.E. Engelstad and O.N. Østerbø, "Analysis of the Total Delay of IEEE 802.11e EDCA and 802.11 DCF," *Proc. IEEE Int'l Conf. Comm. (ICC '06),* vol. 2, pp. 552-559, 2006.
[14] J.-D. Kim and C.-K. Kim, "Performance Analysis and Evaluation of IEEE 802.11e EDCF," *Wireless Comm. and Mobile Computing,* vol. 4, no. 1, pp. 55-74, 2004.
[15] J.W. Robinson and T.S. Randhawa, "Saturation Throughput Analysis of IEEE 802.11e Enhanced Distributed Coordination Function," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 5, pp. 917-928, 2004.
[16] J.W. Robinson and T.S. Randhawa, "A Practical Model for Transmission Delay of IEEE 802.11e Enhanced Distributed Channel Access," *Proc. IEEE Int'l Personal, Indoor and Mobile Radio Comm. (PIMRC '04),* vol. 1, pp. 323-328, 2004.
[17] F. Peng, H.M. Alnuweiri, and V.C.M. Leung, "Analysis of Burst Transmission in IEEE 802.11e Wireless LANs," *Proc. IEEE Int'l Conf. Comm. (ICC '05),* vol. 2, pp. 535-539, 2005.
[18] Y. Ge, J.C. Hou, and S. Choi, "An Analytic Study of Tuning Systems Parameters in IEEE 802.11e Enhanced Distributed Channel Access," *Computer Networks,* vol. 51, no. 8, pp. 1955-1980, 2007.
[19] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE J. Selected Areas in Comm.,* vol. 18, no. 3, pp. 535-547, 2000.
[20] V. Ramaiyan, A. Kumar, and E. Altman, "Fixed Point Analysis of Single Cell IEEE 802.11e WLANs: Uniqueness, Multistability and Throughput Differentiation," *Proc. ACM SIGMETRICS,* 2005.
[21] D. Xu, T. Sakurai, and H.L. Vu, "An Access Delay Model for IEEE 802.11e EDCA," technical report, Dept. of Electrical and Electronic Eng., Univ. of Melbourne, http://www.cubinlab.ee.unimelb.edu.au/~xud, 2007.
[22] B.-J. Kwak, N.-O. Song, and L. Miller, "Performance Analysis of Exponential Backoff," *IEEE/ACM Trans. Networking,* vol. 13, no. 2, pp. 343-355, 2005.
[23] J. Abate, G.L. Choudhury, and W. Whitt, "An Introduction to Numerical Transform Inversion and Its Application to Probability Models," *Computational Probability,* W.K. Grassman, ed., pp. 257-323, Kluwer Academic Publishers, 2000.
[24] *The Network Simulator ns-2,* http://www.isi.edu/nsnam/ns/, 2008.
[25] S. Wiethölter and C. Hoene, *An IEEE 802.11e EDCF and CFB Simulation Model for ns-2,* http://www.tkn.tu-berlinde/research/802.11e_ns2/, 2008.

Dongxia Xu received the BE degree from the University of Electronic Science and Technology of China in 1999. She is currently a PhD candidate in the Department of Electrical and Electronic Engineering, University of Melbourne. From 1999 to 2004, she was a technical support engineer at Lucent Technologies, China. Her research interests include performance analysis of MAC protocols in WLANs. She is a student member of the IEEE.

Taka Sakurai received the BSc degree in applied mathematics and the BE degree in electrical engineering from the University of Adelaide in 1988 and 1989, respectively, and the PhD 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. He is currently with the Chief Technology Office of Telstra Corp. and an honorary fellow of the University of Melbourne. His research interests include the areas of performance analysis of wireless networks, design of MAC protocols for wireless LANS and sensor networks, and computational probability. He is a member of the IEEE.

Hai L. Vu received the MSc and PhD degrees in electrical engineering from the Technical University of Budapest 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. From 2000 to 2005, he was with the Department of Electrical and Electronic Engineering, University of Melbourne. In 2005, he joined the Swinburne University of Technology, Hawthorn, Victoria, Australia, where he is currently with the Centre for Advanced Internet Architectures (CAIA), Faculty of Information and Communication Technologies (FICT). He has authored or coauthored more than 70 scientific journals and conference papers. His current research interests include data network modeling, the performance evaluation of wireless and optical networks, and telecommunication networks design. He is a senior member of the IEEE.