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*— In this paper, we analyse the MAC access delay of the IEEE 802.11e EDCA mechanism under saturation. We develop a detailed analytical model to evaluate the influence of all EDCA differentiation parameters, namely AIFS, CWmin, CWmax and $\mathrm{TXOP\ limit}$, as well as the backoff multiplier β. We derive explicit expressions for the mean, standard deviation and generating function of the access delay distribution. By applying numerical inversion on the generating function, we are able to efficiently compute values of the distribution. Through comparison with simulation, we confirm the accuracy of our analytical model over a wide range of operating conditions. Using the model, 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 study the characteristics of CWmin, AIFS, TXOP, and β differentiation. We find that, though rejected during the standardization process, β differentiation 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.**

I. 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 ($\mathrm{TXOP\ limit}$). A fifth parameter representing the backoff window multiplier (sometimes called the persistence factor), 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 the present 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 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. Note that parts of this work have appeared previously in conference form [5], [6].

Many recent papers have proposed analytical models for various subsets of EDCA functionality [7]-[19]. Xiao [7] models CWmin and CWmax differentiation, [8]-[18] model CWmin, CWmax and AIFS differentiation, and Peng et. al. [19] develop a simple model for TXOP differentiation only. Compared to previous models, our model is novel for the following reasons: (i) it correctly accounts for all 4 differentiation parameters in the standard; (ii) it yields the standard deviation and distributional values of the access delay, as well as the commonly obtained mean access delay; (iii) and it provides accurate estimates of these metrics. Ge at al. [20] attemp to explicitly account for all differentiation parameters in their model, but they actually analyse and simulate a p-persistent version of EDCA, which does not have the same characteristics as EDCA. In [14], it is stated that a 4-parameter model can be built by simply inflating the packet length in their 3-parameter model to account for TXOP differentiation. However, as we will show in our model development, an accurate model of TXOP differentiation is a non-trivial 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 a fully integrated one that can capture joint differentiation by up to four parameters (or five parameters including β). However, for ease of understanding, we present the model in terms of three sub-models: a collision probability model that estimates the collision probabilities of the different classes; a delay model that accounts for all phenomena 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 modelling extension.

A collision probability model is a vital element of any

D. Xu is with the National ICT Australia (NICTA), Victoria Laboratory, Department of Electrical and Electronic Engineering, The University of Melbourne, VIC 3010, Australia.

T. Sakurai is with the Department of Electrical and Electronic Engineering, The University of Melbourne, VIC 3010, Australia.

H. L. Vu is with the Centre for Advanced Internet Architectures, Faculty of I.C.T., Swinburne Univ. of Technology, P.O. Box 218, VIC 3122, Australia.

EDCA analysis. All the aforementioned studies use extensions of Bianchi's two-dimensional (2-D) Markov chain analysis of DCF [21] to derive the collision probabilities, though [12] shows that there are other approaches. To incorporate AIFS differentiation, [8]-[10] resort to 3-D Markov chains, while [11] uses a 4-D Markov chain. In contrast, [16] and [17] develop less complex models based on separate 2- and 1-D Markov chains. Our collision probability model is based on that of [16], but uses an average value analysis in place of the 2-D 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 analyse only throughput and/or mean delay. Exceptions are [10], where the delay distribution is obtained using a computational approach based on the transient analysis of a Markov chain; [18], where the delay distribution is approximated by estimating the probabilities of alternate delay outcomes; and Engelstad and Østerbø [15], where points of the distribution are obtained by inverting the generating function of the delay distribution. In our study, we present a more direct and accurate method to obtain the delay distribution. Similar to [15], we derive the generating function of the distribution of the access delay and obtain distributional values via numerical transform inversion. However, our generating function is more detailed and accurate than that of [15], 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 and not by differentiation and limit-taking of the generating function, which is the approach used in [15]. The direct approach is advantageous because the generating function in question is complicated, making differentiation tedious. Perhaps as a result of this complexity, Engelstad and Østerbø [15] 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 use to develop analytical insights into the relative importance of parameters and to quantify the jitter performance of the differentiation mechanisms.

Achieving accuracy in the distributional values clearly demands a more detailed analysis than one that is sufficient for delivering accuracy in throughput or mean delay. In our delay model, we carefully account for all events that noticeably contribute to the access delay of a packet from a tagged station. We include the delays due to the backoff process of the tagged station, interruptions to the countdown of the AIFS guard-time by higher priority stations, collisions involving the tagged station, and transmissions and collisions involving other stations. We develop the delay and TXOP models in terms of random variables, which makes it possible to readily obtain explicit expressions for the mean, standard deviation, and generating function. 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. Significantly, we have found that our analytical tail distribution is typically an excellent match with simulation down to 10^{-3}, and often beyond.

In addition to developing an analytical 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 contending 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 contending stations. The asymptotics and approximations reveal the salient model parameters for performance under different differentiation mechanisms, 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. [22], who obtained asymptotics for throughput ratios under CWmin, AIFS and β only 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 [22], we also derive asymptotic results for the individual ACs rather than the ratios, as well as 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 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. We also find that CWmin and AIFS individually provide only coarse-grained differentiation, but that joint CWmin and AIFS can provide access to intermediate differentiation levels.

The rest of this paper is organized as follows. In Section II, we give a summary of the EDCA mechanism. As EDCA has been thoroughly reviewed in many previous papers (e.g. [23]), we keep our description brief. In Section III, we present our analytical model for the MAC access delay, starting with the collision probability model. Then we describe our access delay model that takes into account the CWmin, CWmax, AIFS and β mechanisms, and derive expressions for the associated mean, standard deviation and generating function. At the end of this section, we present our TXOP model, and derive the mean, standard deviation and generating function when all five differentiation parameters are included. In Section IV, we present asymptotics and approximations for the mean and standard deviation. The validation of the analytical model with ns-2 simulation is carried out in Section V, 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 VI.

II. Overview of EDCA

EDCA is a prioritized carrier sense multiple access with collision avoidance (CSMA/CA) access mechanism which uses (truncated) binary exponential backoff (BEB). It realizes service differentiation through the use of four ACs in each station. Each AC has its own transmission queue and four adjustable contention parameters: CWmin, CWmax, AIFS and $\mathrm{TXOP\ limit}$. When a packet arrives at the MAC layer from the higher layers, it is assigned to one of the ACs according to its user priority. The parameter values of different ACs should differ in at least one parameter to enable differentiation.

The CWmin and CWmax parameters define the initial and maximum values of the contention window (CW) used in the backoff process. In this process, a discrete backoff time measured in backoff slots is randomly selected from [0,CW-1]. A backoff entity is maintained by each AC in the station. The backoff timer counts down as long as the channel is idle but is frozen when the channel is busy. When the backoff timer reaches zero, the station starts transmitting. If the transmission is successful, the receiving MAC layer sends an ACK (acknowledgement) after a short interframe spacing time, SIFS. Upon failure to receive an ACK (indicating an errored transmission or collision), the CWs of the senders are doubled, and the packets are scheduled for retransmission. Doubling of CW continues in response to further collisions until CWmax is reached, after which CW is maintained at CWmax until the packet is successfully transmitted, or until the maximum permitted number of attempts is reached.

The AIFS parameter defines the guard time that a station must observe after a busy channel period before its backoff timer can be resumed. A smaller AIFS means a higher priority of access. The value of AIFS is always greater than SIFS to ensure contention-free access for ACKs and other control packets. If an AIFS countdown is interrupted by a transmission from a higher priority station, the countdown is stopped and a new AIFS countdown is started when the channel becomes idle.

The $\mathrm{TXOP\ limit}$ parameter defines the maximum duration for which a station can enjoy uninterrupted control of the medium after obtaining a transmission opportunity. Uninterrupted control is guaranteed by allowing the station to send its next data packet after a SIFS time following the receipt of an ACK for the previous packet. A value of $\mathrm{TXOP\ limit} = 0$ indicates only a single packet may be transmitted for each transmission opportunity.

Like DCF, EDCA can operate in either two-way (DATA-ACK) or four-way (RTS-CTS-DATA-ACK) handshaking modes. In our analysis, we cover the two-way handshaking mode only, but the analysis can be readily extended to the four-way mode.

III. ANALYTICAL MODEL

In our model, we make the following assumptions:(i) all stations are saturated (always have a packet to send);(ii) the collision probability is constant regardless of the state, but may differ with AC; (iii) channel conditions are ideal; (iv) ACK packets are transmitted at the lowest basic rate and the ACK timeout after a collision matches the guard time observed by non-colliding nodes, and (v) each station only has traffic belonging to a single AC. The first four assumptions are standard for studies of 802.11 performance and originate from [21]. Assumptions (iv) and (v) can be removed at the expense of additional modelling 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 non-decreasing 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 $\beta_k > 1$, instead of binary exponential backoff as in the standard. The maximum backoff stage for $\mathrm{AC}[k]$ is m_k, so that the maximum contention window is $\mathrm{CWmax}_k = \langle \beta_k^{m_k} \mathrm{W}_k \rangle$, where $\langle . \rangle$ denotes rounding to the nearest integer. The transmission opportunity limit for $\mathrm{AC}[k]$ is denoted by TXOP_k.

A. 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 collision probabilities expressed in terms of the transmission probabilities, with an opposing set of equations for the transmission probabilities expressed in terms of the collision probabilities. We obtain the former set of equations by following an approach proposed by Kim and Kim [16], which we summarize below.

Kim and Kim [16], and also Robinson and Randhawa [17], use the concept of *slot class* 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 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.

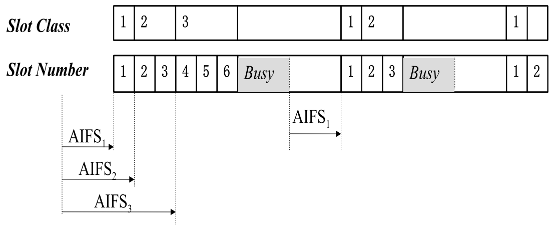

Fig. 1. Slot Number and Slot Class

In slot class j, only stations with access category $k \leq j$ can transmit. This gives rise to the notion of a conditional collision probability $c_k(j)$ for 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 define $r_i = 1 - p_i$.

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 [16], 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

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

Equations (1), (2) and (3) express c_k as a non-linear function of the transmission probabilities p_k. To find p_k as a function of the collision probabilities c_k, [16] and [17] use variants of the 2-D Markov chain of [21]. 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 analyse the dynamics of the backoff process in a similar way to Kwak et al. [24], who analysed 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 non-zero transition probabilities

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

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

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

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

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

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

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

where $\eta_k = (1 - c_k)(1 - c_k^R)^{-1}$.

Equations (1), (2), (3), (4) and (7) constitute a non-linear system of equations that can be solved iteratively to obtain the p_k's and c_k's.

B. 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 conditions. From the protocol description in Section II, we can identify several events that contribute to the access delay. The most obvious is simply 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 can be written as

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

where $\epsilon^{(k)}$ is a 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 a 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 non-tagged 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 $\text{TXOP}_i = 0$ ($i = 1, \ldots, J$), 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. Later in Section III-E, 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 III-A, 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

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

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

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

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

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

where w.p. stands for 'with probability'; T^* is the channel occupancy of a successful transmission from a higher priority station; 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 [1]

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

and

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

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 interpreted 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 countdown of AIFS_k 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, we have

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

where $i_1, i_2, \ldots, i_{h^{(k)}} = 0, 1, \ldots \infty$ are non-negative integers. The integers $i_1, i_2, \ldots, i_{h^{(k)}}$ 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). It can be confirmed that the probabilities in (17) sum to one through an application of the multinomial theorem.

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 $i = 0, ..., R-1$. We can therefore write

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

where $i = 0, ..., R-1$. 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

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

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 and fixed packet lengths, we have

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

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 non-tagged station, or by a collision involving the non-tagged 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

$$B_j^{(k)} = \sum_{n=1}^{U_j^{(k)}} Y_n^{(k)}, \tag{22}$$

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 non-tagged station attempts to transmit, $Y^{(k)}$ equals the channel occupancy of a collision involving non-tagged stations, denoted by $H^{(k)}$. Hence we obtain

$$Y^{(k)} = \begin{cases} t_{slot} & w.p. \quad 1 - c_k \\ G^{(k)} & w.p. \quad \gamma^{(k)} \\ H^{(k)} & w.p. \quad \nu^{(k)}, \end{cases} \tag{23}$$

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 the different slot classes:

$$\gamma^{(k)} \quad = \quad \sum_{j=k}^{J} \gamma^{(k)}(j) \frac{P(j)}{\sum_{i=k}^{J} P(i)},$$

where $\gamma^{(k)}(j)$ can be obtained as (21). The first term in (21) is the probability that exactly one of the non-tagged 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. Given $\gamma^{(k)}$, $\nu^{(k)}$ can be computed from $\nu^{(k)} = c_k - \gamma^{(k)}$. In the case of uniform, fixed length packets, we have

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

C. *Generating Function*

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

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

All the r.v.'s introduced in III-B are non-negative, but not always integer-valued. However, they can be easily transformed to integer-valued r.v's by defining a lattice with 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 III-B to refer to their integer-valued equivalents. For example, we write $P(D^{(k)} = r), r = 0, 1, \dots$ for the pmf of the integer-valued access delay $D^{(k)}$, and $\widehat{D^{(k)}}(z)$ for the generating function.

We can immediately obtain an expression for $\widehat{D^{(k)}}(z)$ from (8):

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

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

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

Based on (18), we can find $\widehat{A}(z)$ as:

$$\widehat{A}(z) = \sum_{i=0}^{R-1} \eta_k c_k^i \widehat{A}_i(z). \tag{27}$$

From (19), we obtain

$$\widehat{A}_i(z) = \widehat{C}(z)^i \prod_{j=0}^{i} \widehat{B}_j(z). \tag{28}$$

It follows from (20) that

$$\widehat{C}(z) = \widehat{\epsilon}(z) z^{\omega}, \tag{29}$$

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

From (22), the generating function of $B_j^{(k)}$ is given by

$$\widehat{B}_j(z) = \widehat{U}_j(\widehat{Y}(z)). \tag{30}$$

Equation (5) yields

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

where $f(j) = \langle \beta_k^j \mathrm{W}_k \rangle$.

From (23) it follows that

$$\widehat{Y}(z) = (1 - c_k) z^{t_{slot}/\delta} + \gamma \widehat{G}(z) + \nu \widehat{H}(z), \tag{31}$$

where it is easy to obtain from (24) that

$$\widehat{G}(z) = \widehat{H}(z) = \widehat{\epsilon}(z) z^{\omega}. \tag{32}$$

The next step is to find the generating function of $\epsilon^{(k)}$. For the highest priority class, AC$[1]$, we have that

$$\widehat{\epsilon}(z) = z^{\text{AIFS}_1/\delta}. \tag{33}$$

$$\gamma^{(k)}(j) \quad = \quad (n_k-1)p_k r_k^{n_k-2}\prod_{\substack{i=1\\ i\neq k}}^{j} r_i^{n_i} + r_k^{n_k-1}\sum_{\substack{i=1\\ i\neq k}}^{j}[n_i p_i r_i^{n_i-1}\prod_{\substack{l=1\\ l\neq k, l\neq i}}^{j} r_l^{n_l}]. \tag{21}$$

$$\begin{aligned}
\mathrm{E}[Y^{(k)}] &= (1-c_k)t_{slot} + \gamma^{(k)}\,\mathrm{E}[G^{(k)}] + \nu^{(k)}\,\mathrm{E}[H^{(k)}], \qquad (34)\\
\mathrm{V}[Y^{(k)}] &= (1-c_k)(t_{slot}-\mathrm{E}[Y^{(k)}])^2 + \gamma^{(k)}(\mathrm{V}[G^{(k)}] + (\mathrm{E}[G^{(k)}]-\mathrm{E}[Y^{(k)}])^2) + \nu^{(k)}(\mathrm{V}[H^{(k)}] + (\mathrm{E}[H^{(k)}]-\mathrm{E}[Y^{(k)}])^2).
\end{aligned}$$

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

$$\widehat{\epsilon}(z) \quad = \quad \frac{z^{\mathrm{AIFS}_k/\delta} s}{1-\sum_{l=1}^{h} z^{t_l/\delta}\mu_l}. \tag{35}$$

For fixed length packets, we find that

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

Thus, the generating function of the pmf of the access delay can be derived from equations (25) - (35).

In the numerical experiments reported in Section V-A, 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

$$\widehat{D_c}(z) = \frac{1-\widehat{D}(z)}{1-z}. \tag{36}$$

The analytical distribution results reported in Section V are obtained by numerically inverting (36). We use the LATTICE-POISSON numerical inversion algorithm developed by Abate et. al. [25].

D. Mean and Standard Deviation

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

$$\begin{aligned}
\mathrm{E}[D^{(k)}] &= \mathrm{E}[\epsilon^{(k)}] + \mathrm{E}[A^{(k)}] + \mathrm{E}[T^{(k)}]\\
\mathrm{S}[D^{(k)}] &= \sqrt{\mathrm{V}[\epsilon^{(k)}] + \mathrm{V}[A^{(k)}] + \mathrm{V}[T^{(k)}]},
\end{aligned}$$

where V[.] denotes the variance.

In the case of fixed length packets, we have

$$\mathrm{E}[T^{(k)}] = t_{data}, \qquad \mathrm{V}[T^{(k)}] = 0.$$

For AC[1], it always holds that

$$\mathrm{E}[\epsilon^{(1)}] = \mathrm{AIFS}_1, \qquad \mathrm{V}[\epsilon^{(1)}] = 0.$$

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

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

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

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

where from (19), we have

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

In the case of uniform, fixed packet lengths, it follows from (20) that

$$\begin{aligned}
\mathrm{E}[C^{(k)}] &= t_{data} + \mathrm{SIFS} + t_{ack} + \mathrm{E}[\epsilon^{(k)}],\\
\mathrm{V}[C^{(k)}] &= \mathrm{V}[\epsilon^{(k)}].
\end{aligned}$$

The mean and variance of $B_j^{(k)}$ can be obtained from (22):

$$\begin{aligned}
\mathrm{E}[B_j^{(k)}] &= \mathrm{E}[U_j^{(k)}]\,\mathrm{E}[Y^{(k)}],\\
\mathrm{V}[B_j^{(k)}] &= \mathrm{E}[U_j^{(k)}]\,\mathrm{V}[Y^{(k)}] + \mathrm{E}[Y^{(k)}]^2\,\mathrm{V}[U_j^{(k)}].
\end{aligned}$$

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

$$\mathrm{V}[U_j^{(k)}] = \begin{cases} \frac{1}{12}(\langle\beta_k^j \mathrm{W}_k\rangle^2 - 1) & \text{for} \quad j = 0, ..., m_k - 1,\\ \frac{1}{12}(\langle\beta_k^{m_k} \mathrm{W}_k\rangle^2 - 1) & \text{for} \quad j = m_k, ..., R-1. \end{cases} \tag{37}$$

It can be seen from (23) that the distribution of $Y^{(k)}$ is a simple mixture, so the mean and variance can be written as in (34). For the case of uniform, fixed length packets we have

$$\begin{aligned}
\mathrm{E}[G^{(k)}] &= \mathrm{E}[H^{(k)}] = t_{data} + \mathrm{SIFS} + t_{ack} + \mathrm{E}[\epsilon^{(k)}],\\
\mathrm{V}[G^{(k)}] &= \mathrm{V}[H^{(k)}] = \mathrm{V}[\epsilon^{(k)}],
\end{aligned}$$

Based on the equations above, the expressions for the mean and the variance of the access delay D can be obtained as in (38) and (39).

$$\mathrm{E}[D^{(k)}] = \eta_k \sum_{i=0}^{R-1} c_k^i \{\mathrm{E}[Y^{(k)}] \sum_{j=0}^{i} \mathrm{E}[U_j^{(k)}] + i\,\mathrm{E}[C^{(k)}]\} + \mathrm{E}[T^{(k)}] + \mathrm{E}[\epsilon^{(k)}], \tag{38}$$

$$\begin{aligned} \mathrm{V}[D^{(k)}] &= \eta_k \sum_{i=0}^{R-1} c_k^i \{\sum_{j=0}^{i} (\mathrm{E}[U_j^{(k)}]\,\mathrm{V}[Y^{(k)}] + \mathrm{E}[Y^{(k)}]^2\,\mathrm{V}[U_j^{(k)}]) \\ &\quad + i\,\mathrm{V}[C^{(k)}] + (\mathrm{E}[Y^{(k)}] \sum_{j=0}^{i} \mathrm{E}[U_j^{(k)}] + i\,\mathrm{E}[C^{(k)}] - \mathrm{E}[A^{(k)}])^2\} + \mathrm{V}[T^{(k)}] + \mathrm{V}[\epsilon^{(k)}]. \end{aligned} \tag{39}$$

E. TXOP *Model*

In this section, we analyse the access delay when differentiation by TXOP is configured. Suppose $\mathrm{TXOP}_k > 0$ 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, and since successive DATA-ACK exchanges are separated only by SIFS intervals, collisions cannot occur except to the first transmitted packet.

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

$$D^{(k)} = \begin{cases} D_1^{(k)} & w.p. \quad 1/N_k \\ D_2^{(k)} & w.p. \quad 1/N_k \\ \dots & \\ D_{N_k}^{(k)} & w.p. \quad 1/N_k, \end{cases} \tag{40}$$

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

$$D_i^{(k)} = \mathrm{SIFS} + t_{data}, \tag{41}$$

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

$$D_1^{(k)} = \epsilon^{(k)} + A^{(k)} + t_{data}, \tag{42}$$

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 equations (9) - (17), but with modifications to the expressions for X_i to separately account for different transmission durations between classes:

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

where T_l^* is the channel occupancy of a successful transmission from an AC$[l]$ station; 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

$$\begin{aligned} T_l^* &= \Delta_l + \mathrm{SIFS} + t_{ack} \\ C^* &= t_{data} + \mathrm{SIFS} + t_{ack}. \end{aligned}$$

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

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

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

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

where the probability of exactly one transmission given by the numerator is conditioned by the probability of at least one transmission in the denominator.

An expression for $A^{(k)}$ can be obtained using equations (18) - (22), together with the following modifications to $Y^{(k)}$ to separately account for different transmission durations between classes:

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

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

$$\begin{aligned} G_l^{(k)} &= \Delta_l + \mathrm{SIFS} + t_{ack} + \epsilon^{(k)} \\ H^{(k)} &= t_{data} + \mathrm{SIFS} + t_{ack} + \epsilon^{(k)}. \end{aligned}$$

The $\nu^{(k)}$ is obtained from $\nu^{(k)} = c_k - \sum_{l=1}^{J} \gamma_l^{(k)}$, and $\gamma_l^{(k)}$ can be determined from the weighted average of conditional probabilities in a similar fashion to the collision probability in Section III-A, namely,

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

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 conditional probabilities $\gamma_l^{(k)}(j)$ are given by

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

From expressions (41) and (42), 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 III-D; for $i=2,3,\dots,N_k$, it follows that

$$\begin{aligned} \mathrm{E}[D_i^{(k)}] &= \mathrm{SIFS}+t_{data}, \\ \mathrm{V}[D_i^{(k)}] &= 0, \\ \widehat{D_i^{(k)}}(z) &= z^{(\mathrm{SIFS}+t_{data})/\delta}. \end{aligned}$$

Finally, the mean, standard deviation and generating function of the pmf of $D^{(k)}$ follow from (40) as:

$$\begin{aligned} \mathrm{E}[D^{(k)}] &= \frac{1}{N_k}\sum_{i=1}^{N_k}\mathrm{E}[D_i^{(k)}] \qquad (43) \\ \mathrm{S}[D^{(k)}] &= \sqrt{\frac{1}{N_k}\sum_{i=1}^{N_k}[\mathrm{V}[D_i^{(k)}]+(\mathrm{E}[D_i^{(k)}]-\mathrm{E}[D^{(k)}])^2]} \\ \widehat{D^{(k)}}(z) &= \frac{1}{N_k}\sum_{i=1}^{N_k}\widehat{D_i^{(k)}}(z). \end{aligned}$$

IV. Asymptotic Analysis and Approximations

The expressions for the delay metrics found in Section III are accurate (as we demonstrate in Section V-A) 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 (37), 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.

A. 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. [22] previously studied asymptotic results for throughput ratios under the same conditions, and we make use of some of their intermediate results.

1) TXOP $=0$: From the expression for the mean delay in (38), when $R=\infty$, we obtain

$$\begin{aligned} \mathrm{E}[D^{(k)}] &= \frac{(1-c_k)t_{slot}+c_k\,\mathrm{E}[C^{(k)}]}{p_k(1-c_k)}+\frac{c_k\,\mathrm{E}[C^{(k)}]}{1-c_k} \\ &\quad +t_{data}+\mathrm{E}[\epsilon^{(k)}]. \qquad (44) \end{aligned}$$

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

i) $\mathrm{CW_{min}}$ 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 Ramaiyan et. al. [22] that when $m=R=\infty$, for $k=1,2$, we have

$$\lim_{n\to\infty}c_k\uparrow\frac{1}{\beta},\qquad \lim_{n\to\infty}p_k\downarrow 0. \qquad (45)$$

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

$$p_k=\frac{1-\beta c_k}{\frac{W_k}{2}(1-c_k)},\quad 0\le c_k<\frac{1}{\beta}. \qquad (46)$$

Taking the limit of θ_m using (44) and applying (45) and (46) leads to the result. ■

ii) AIFS differentiation

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

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

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

Proof: In [22], it is shown that for AIFS differentiation, when $m=R=\infty$, (45) still holds, and, in addition,

$$\begin{aligned} \lim_{n\to\infty}n_1p_1 &\uparrow \ln\frac{\beta}{\beta-1}, \\ \lim_{n\to\infty}n_2p_2 &= 0. \qquad (47) \end{aligned}$$

Taking the limit of $\mathrm{E}[D^{(1)}]$ using (44) and applying (45) and (47) 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. ■

iii) β differentiation

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

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

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 [22]:

$$\begin{aligned} \lim_{n\to\infty}c_1 &\uparrow \frac{1}{\beta_1}\text{ and }\lim_{n\to\infty}c_2\uparrow\frac{1}{\beta_1}, \\ \lim_{n\to\infty}n_1p_1 &\uparrow \ln\frac{\beta_1}{\beta_1-1}, \\ \lim_{n\to\infty}n_2p_2 &= 0. \end{aligned}$$

■

We see that when $m = R = \infty$ and n is large, the asymptotic mean delay ratio 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 differentiations, however, the asymptotic mean delay ratio does not exist, since the mean delay of AC[2] stations can be arbitrary large when $n \to \infty$. On the other hand, the mean delay of AC[1] stations does converge and is given in Lemmas 2 and 3. Ramaiyan et. al. [22] obtained similar asymptotic results for the throughput ratio, but did not provide the asymptotic results for the high priority class.

2) TXOP *differentiation:*

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

Proof: As all the parameters except TXOP limit are identical for the two classes, when $m = R = \infty$ we have

$$\begin{aligned} p_1 = p_2 = p \quad &\text{and} \quad \lim_{n\to\infty} p \downarrow 0, \\ c_1 = c_2 = c \quad &\text{and} \quad \lim_{n\to\infty} c \uparrow \frac{1}{\beta}, \\ \lim_{n\to\infty} np \quad &\uparrow \quad \ln \frac{\beta}{\beta - 1}. \end{aligned} \tag{48}$$

From (43), (44) and (48), we have

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

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

As stated in Theorem 1, the asymptotic mean delay ratio under TXOP differentiation is very simple and depends only on the value of the TXOP limit parameters. Although simple, this result has not been observed previously in the literature.

B. Approximations

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

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

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

1) TXOP $= 0$*:* In this section, we consider differentiation by some or all of CWmin, AIFS and β. Under the assumptions listed previously, we obtain the following approximations:

$$\mathrm{E}[D] \approx \frac{c\beta\Gamma}{p(\beta-1)q^h}, \tag{49}$$

$$\mathrm{V}[D] \approx \frac{c^2W^2\Gamma^2}{q^{2h}} \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \tag{50}$$

where $\Gamma = \mathrm{AIFS}_1 + t_{data} + \mathrm{SIFS} + t_{ack}$, $q = r_1^{n_1}$ and $h = 0$ for class 1 and $h = (\mathrm{AIFS}_2 - \mathrm{AIFS}_1)/t_{slot}$ for class 2. The derivations of (49) and (50) are given in the appendix.

Straightforwardly, the moment ratios are given by

$$\begin{aligned} \theta_m &\approx \frac{p_1c_2}{p_2c_1q^h}\frac{\beta_2(\beta_1-1)}{\beta_1(\beta_2-1)} \\ &\approx \frac{\beta_2(\beta_1-1)}{\beta_1(\beta_2-1)} \frac{W_2c_2\sum_{i=0}^{R-1}\eta_2(\beta_2c_2)^i}{q^hW_1c_1\sum_{i=0}^{R-1}\eta_1(\beta_1c_1)^i}. \end{aligned} \tag{51}$$

$$\begin{aligned} \theta_s &\approx \frac{c_2W_2\beta_2(\beta_1-1)}{c_1W_1\beta_1(\beta_2-1)q^h} \times \\ &\quad \sqrt{\frac{(2\beta_2+1)(\beta_1+1)}{(2\beta_1+1)(\beta_2+1)} \frac{\sum_{i=0}^{R-1}\eta_2c_2^i\beta_2^{2i}}{\sum_{i=0}^{R-1}\eta_1c_1^i\beta_1^{2i}}}. \end{aligned} \tag{52}$$

For CWmin only differentiation, $c_1 \approx c_2$ for sufficiently large n_1 and n_2 [5], so both (51) and (52) simplify to W_2/W_1.

2) TXOP *differentiation:* In this section, we present approximations for TXOP only differentiation. From Section III-A, 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:

$$\mathrm{E}[D] \approx \frac{[c + (n_1N_1 + n_2N_2 - n)p(1-p)^n]\beta\Gamma}{Np(\beta-1)} \tag{53}$$

$$\begin{aligned} \mathrm{V}[D] &\approx \frac{[c + (n_1N_1 + n_2N_2 - n)p(1-p)^n]^2W^2\Gamma^2}{N} \\ &\quad \times \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2} \sum_{i=0}^{R-1} \eta c^i \beta^{2i}, \end{aligned} \tag{54}$$

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

$$\theta_m \approx \frac{N_1}{N_2}, \qquad \theta_s \approx \sqrt{\frac{N_1}{N_2}}. \tag{55}$$

V. Results and Discussion

This section has three objectives: (i) to compare numerical results obtained from the analysis of Section III with simulation in order to confirm the accuracy of the model; (ii) to utilize the model to study the effectiveness of the various differentiation mechanisms for service separation; and (iii) to test the accuracy of the approximations presented in Section IV-B. The simulations were conducted using the ns-2 (version 2.28) simulator [26], combined with an EDCA module developed by TU-Berlin [27]. 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 1000 bytes. The MAC and physical layer parameters were configured in accordance with the default values in IEEE 802.11b, as shown in Table I.

TABLE I
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

Accordingly, the durations for data and acknowledgement packet transmissions used in our analytical model are given by

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

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.

A. Validation

To corroborate the accuracy of the analysis of Section III, 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% 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 a retransmission limit $R = 7$. For all but the last example, the β parameter was maintained at 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 II lists the $n_1 : n_2$ ratios and the differentiation parameters of four scenarios that were investigated. The first three scenarios test the differentiation achieved through only one parameter at a time, namely CWmin, AIFS, and $\mathrm{TXOP\ limit}$, respectively.

TABLE II
EDCA PARAMETERS USED IN SIMULATION

Scen.	$n_1 : n_2$	W		CWmax		AIFS (μs)		TXOP (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	32	32	1024	1024	50	50	2.906	0
4	1 : 1	8	16	512	1024	50	70	2.906	0

The results (mean, standard deviation and CCDF) for scenarios 1 and 2 are shown together in Figs. 2, 3 and 4. The mean and standard deviation results are plotted against different total numbers of stations $n_1 + n_2$, while the CCDF results 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 the third scenario, we evaluate the accuracy of the model when TXOP differentiation is enabled. By setting $\mathrm{TXOP}_1 = 2.906 ms$, a high priority station is allowed to send two consecutive packets once it has obtained access to the channel. The corresponding mean delay, standard deviation and CCDF are plotted in Figs. 5, 6 and 7, respectively. The analytical results are again in excellent agreement with the simulation results.

In Figs. 8, 9 and 10, we present results for the mean, standard deviation and CCDF for the last scenario of Table II. These results demonstrate that our model accurately predicts performance when all four differentiation mechanisms in the standard are activated, namely CWmin, CWmax, AIFS and $\mathrm{TXOP\ limit}$. 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 Figs. 11 and 12 for an example with four ACs. The following parameters settings were used: $\mathrm{W}_k = \{8/8/32/32\}$ and $\mathrm{AIFS}_k = \{50/70/70/90\}\ \mu s$. The values of other parameter were common for all classes: $\mathrm{CWmax} = 1024$ and $\mathrm{TXOP} = 0$.

In terms of a method for obtaining values of the distribution, the generating function analysis of Engelstad and Østerbø [15] comes closest in spirit to our approach. In Fig. 13, we plot the CCDFs obtained by numerically inverting our generating function and inverting the generating function derived in [15] for the saturation condition. The parameters were the same as in scenario 2 of Table II, except that the AIFS of the lower priority stations was set to $90\mu s$. 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 [15] 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 [14], and do not include the additional delay caused by multiple interruptions to the AIFS of low priority stations.

Finally, we present results in Figs. 14 and 15 to confirm that our model correctly predicts performance under β differentiation. We used $\beta_1 = 1.8$ and $\beta_2 = 2$, and all other parameters were the same for both classes: $\mathrm{W} = 32$, $\mathrm{AIFS} = 50\ \mu s$, $\mathrm{CWmax} = 1024$ and $\mathrm{TXOP} = 0$.

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 [22], which results in multiple solutions to the fixed-point.

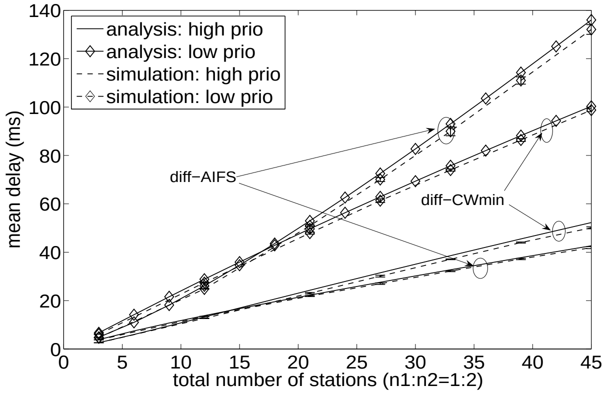

Fig. 2. Mean access delay: differentiation by CWmin or AIFS.

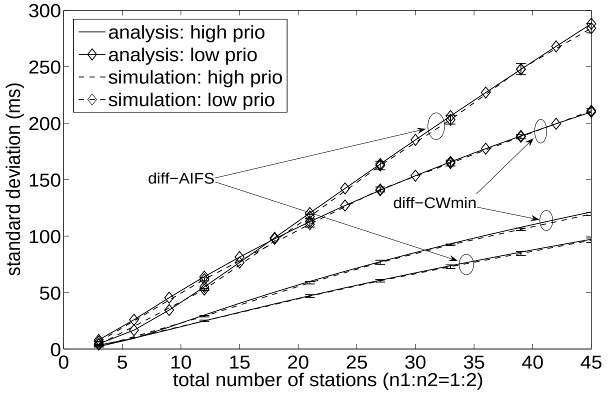

Fig. 3. Standard deviation of access delay: differentiation by CWmin or AIFS.

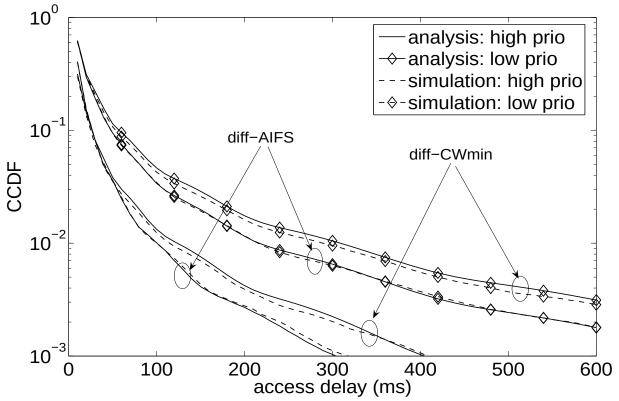

Fig. 4. CCDF of access delay: differentiation by CWmin or AIFS, $n_1 = 4$ and $n_2 = 8$.

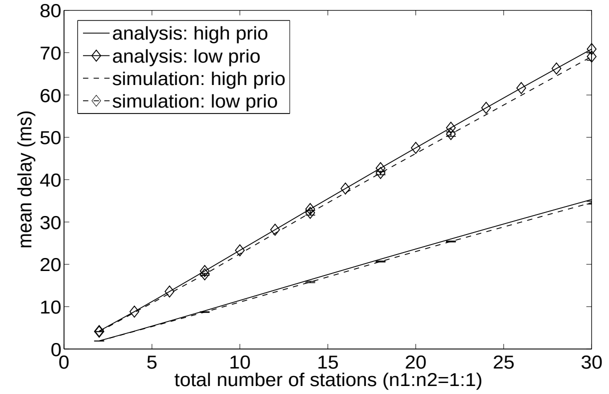

Fig. 5. Mean access delay: differentiation by TXOP.

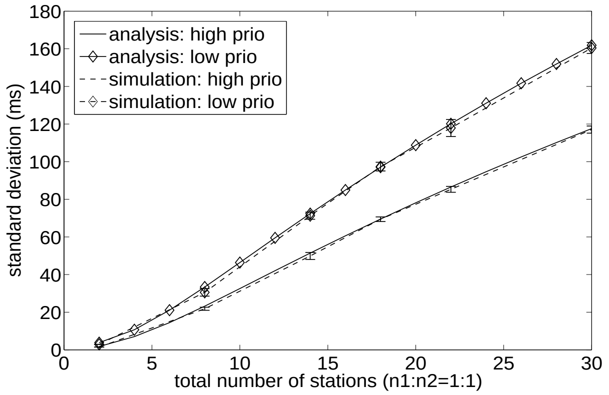

Fig. 6. Standard deviation of access delay: differentiation by TXOP.

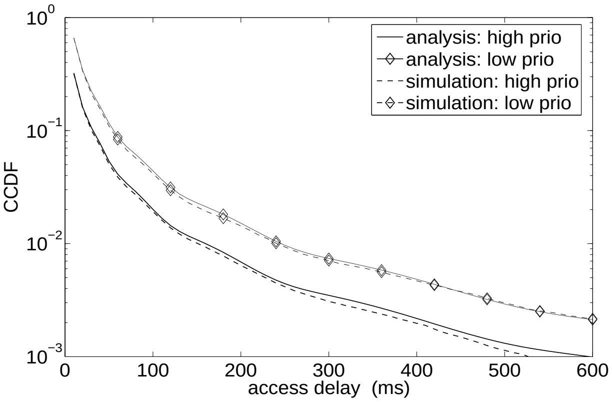

Fig. 7. CCDF of access delay: differentiation by TXOP, $n_1 = 6$ and $n_2 = 6$.

B. 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 IV-B. 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 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 by-product 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.

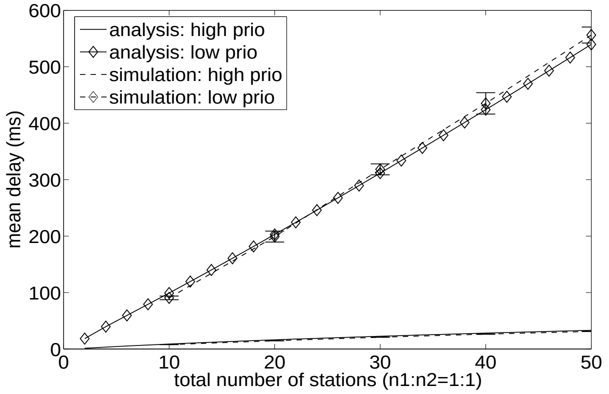

Fig. 8. Mean access delay: differentiation by CWmin, CWmax, AIFS and TXOP.

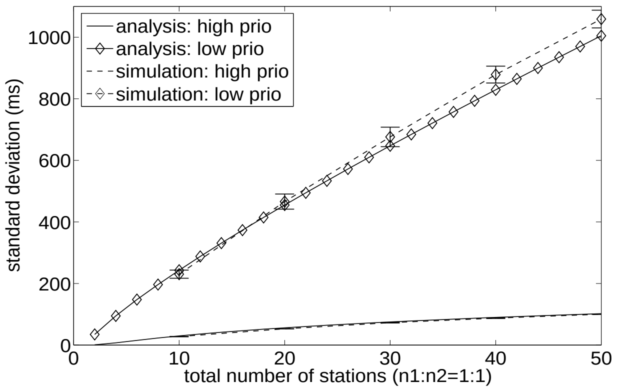

Fig. 9. Standard deviation of access delay: differentiation by CWmin, CWmax, AIFS and TXOP.

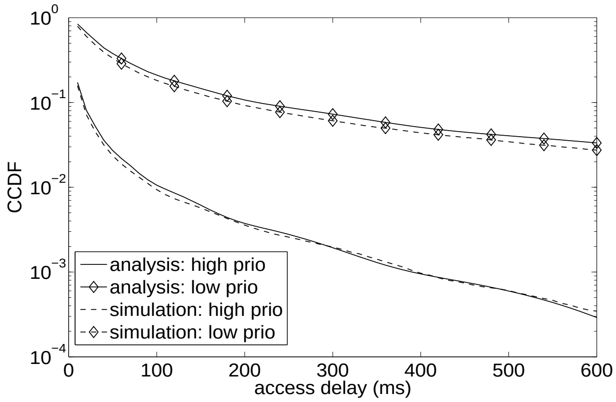

Fig. 10. CCDF of access delay: differentiation by CWmin, CWmax, AIFS and TXOP, $n_1 = n_2 = 5$.

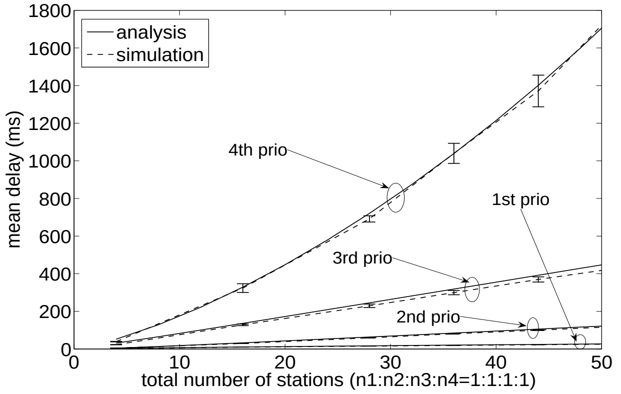

Fig. 11. Mean access delay: four classes.

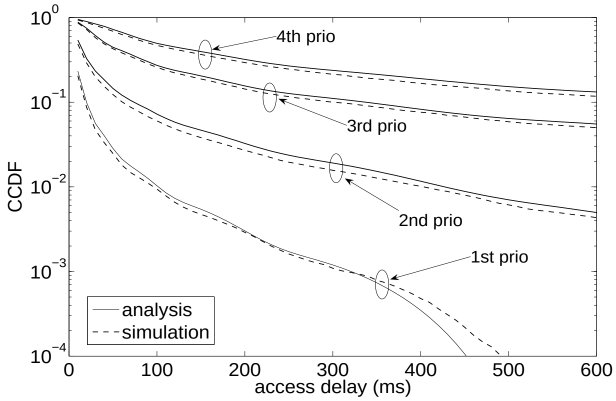

Fig. 12. CCDF of access delay: four classes, $n_1 = n_2 = n_3 = n_4 = 4$

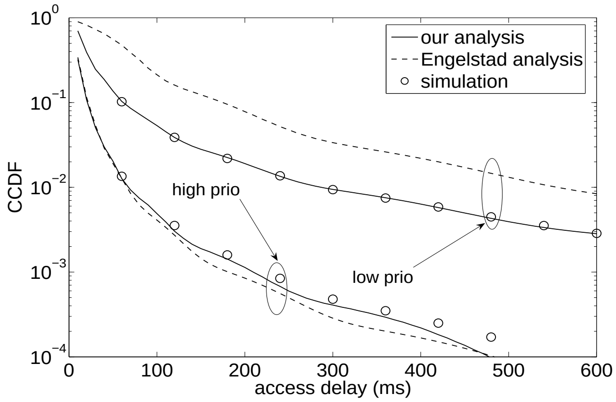

Fig. 13. CCDF of access delay: comparison with Engelstad result, $n_1 = 4$ and $n_2 = 8$.

Consider a setting with two ACs with equal numbers of stations, and define the following reference set of parameter values: $\{\mathrm{W}, \mathrm{AIFS}, \mathrm{TXOP}, \beta\} = \{16, 50\mu s, 0, 2\}$. 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 (55) for TXOP differentiation, and (51) and (52) for the other mechanisms.

The analytical and approximate moment ratios under CWmin differentiation are illustrated in Figs. 16 and 17. 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

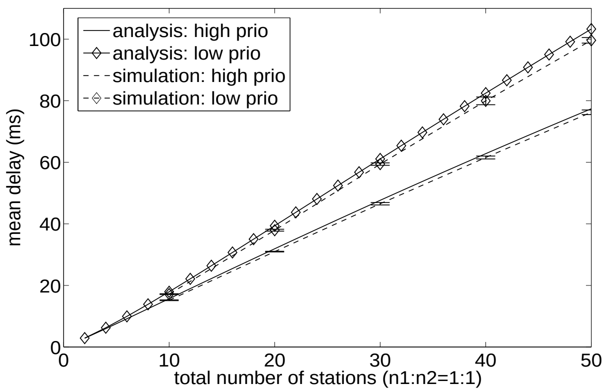

Fig. 14. Mean access delay: differentiation by β.

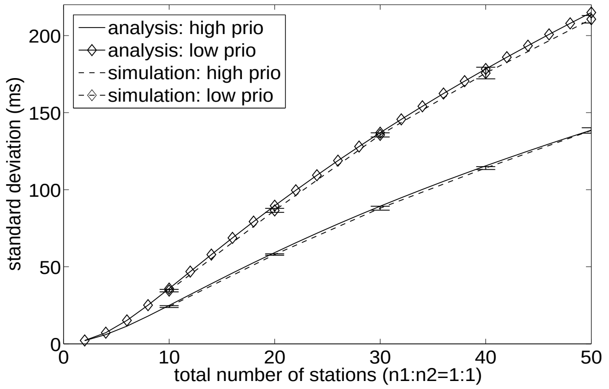

Fig. 15. Standard deviation of access delay: differentiation by β.

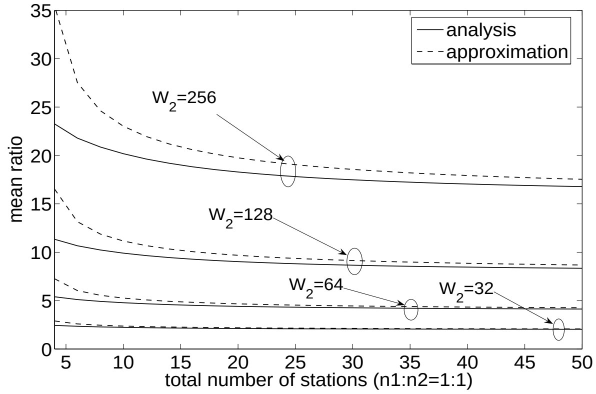

Fig. 16. Mean ratio for $\mathrm{CW}_{\min}$ differentiation.

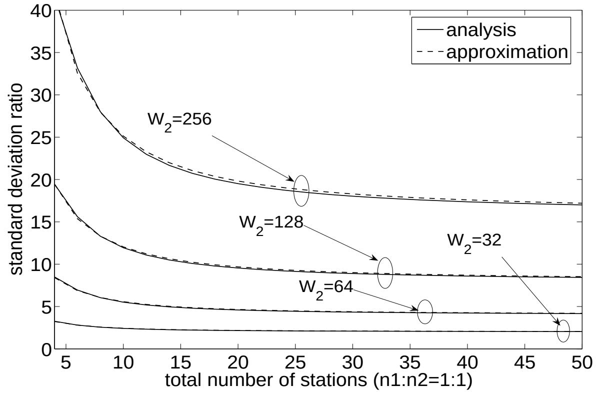

Fig. 17. Standard deviation ratio for $\mathrm{CW}_{\min}$ differentiation.

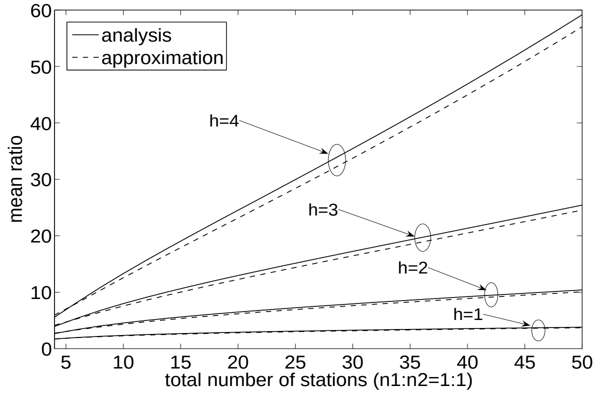

Fig. 18. Mean ratio for AIFS differentiation.

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 IV-A and the observations made in Section IV-B. A consequence of a non-increasing 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. 18 and 19 depict moment ratios when increasing levels of AIFS differentiation are applied. Specifically, AIFS_1 is maintained at $50\mu s$, while $\mathrm{AIFS}_2 = 70, 90, 110, 130\mu s$. 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. 20 and 21, where TXOP_2 is fixed at 0 and TXOP_1 is varied to permit the transmission of $2, 3, 4$ or 5 packets. The shape 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. 22 and 23 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. 16 and 17 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. 16–23 reveal that the approximations are accurate

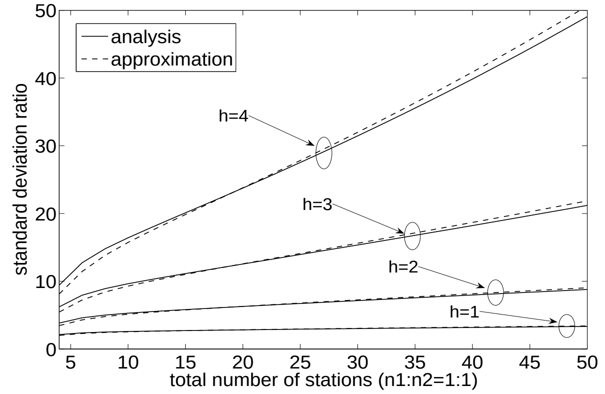

Fig. 19. Standard deviation ratio for AIFS differentiation.

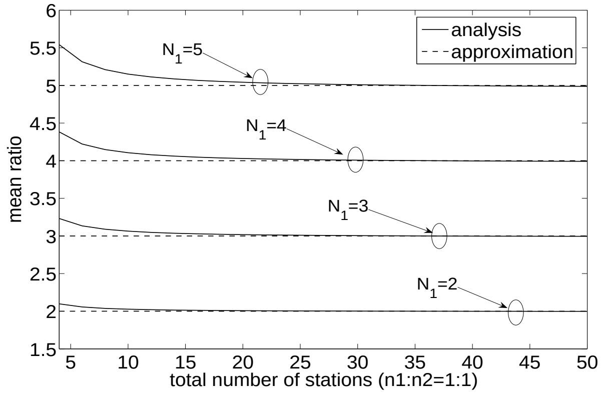

Fig. 20. Mean ratio for TXOP differentiation.

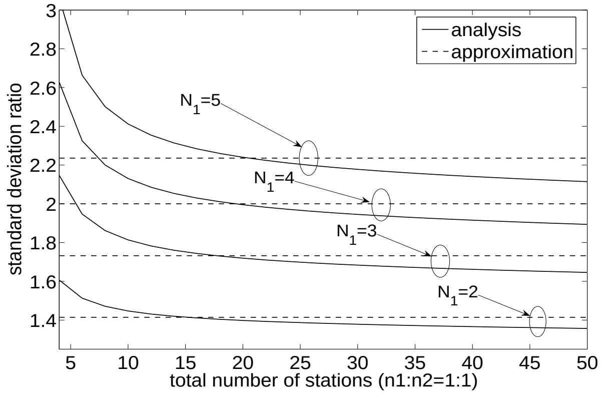

Fig. 21. Standard deviation ratio for TXOP differentiation.

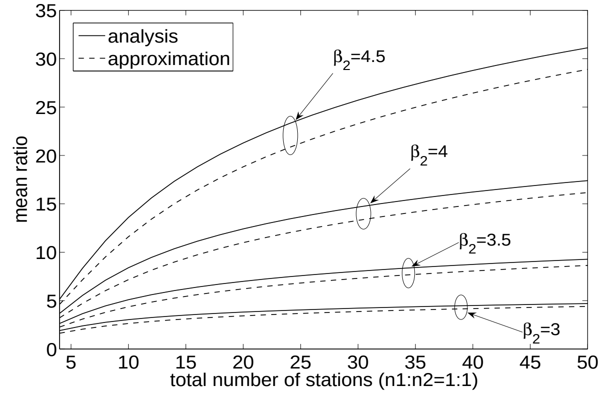

Fig. 22. Mean ratio for β differentiation.

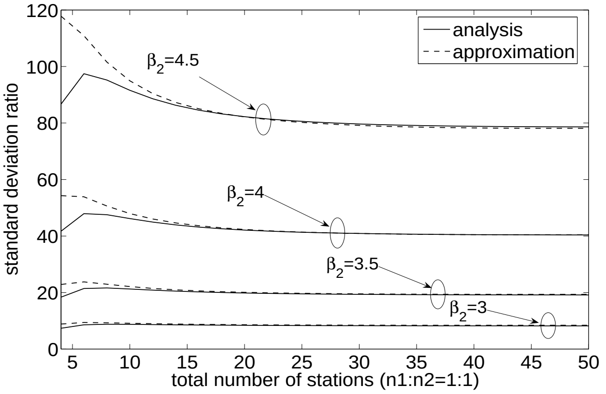

Fig. 23. Standard deviation ratio for β differentiation.

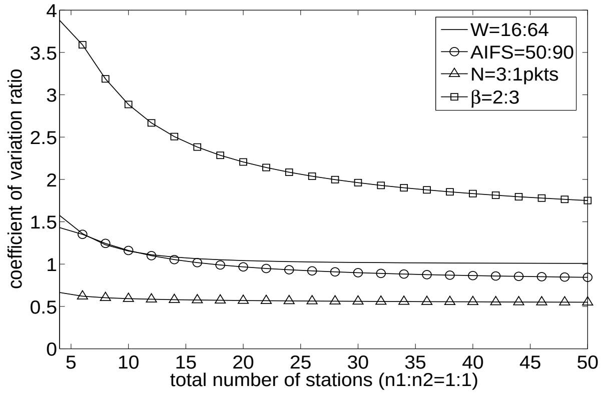

Fig. 24. Coefficient of variation ratio for different parameters.

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 make 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. 24, 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

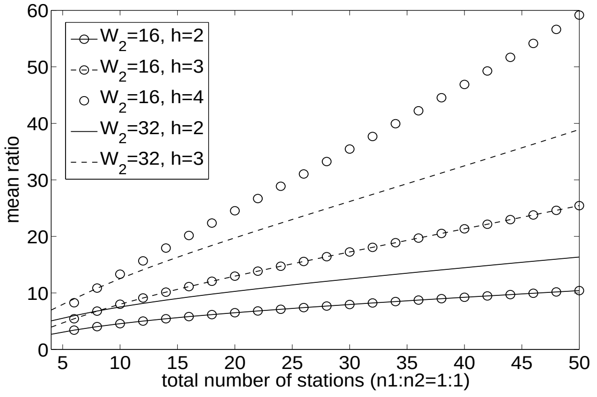

Fig. 25. Mean ratio for joint AIFS and $\mathrm{CW_{min}}$ differentiation.

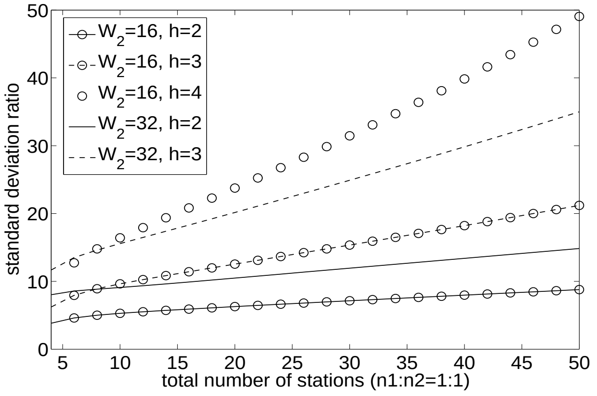

Fig. 26. Standard deviation ratio for joint AIFS and $\mathrm{CW_{min}}$ differentiation.

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.

In [20], it is suggested that to minimize the dimensions of the design problem, single parameter differentiation should be preferred over joint differentiation by multiple parameters. By way of example, it is shown in [20] that differentiation using CWmin, where CWmin can assume any integer value, gives a flexible form of differentiation. However, as demonstrated in Figs. 16 and 17, constraining CWmin to powers of 2 as in the standard limits the differentiation levels that can be achieved. Differentiation by AIFS is also coarse-grained, as shown in Figs. 18 and 19. To achieve intermediate levels of differentiation, it may be necessary to resort to joint differentiation. Figs. 25 and 26 show curves for mean and standard deviation ratios for AIFS only differentiation and differentiation by AIFS and CWmin in unison. Keeping all other parameters settings at the reference values, AIFS_2 and W_2 are varied as indicated in the legend (the curves with $W_2 = 16$ correspond to differentiation by AIFS only). The results confirm that joint differentiation is a useful means of creating intermediate differentiation levels.

VI. 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. Explicit expressions for the mean, standard deviation and generating function of the distribution of the access delay were obtained. The model captures all four tunable parameters defined in the EDCA standard, namely CWmin, CWmax, AIFS and $\mathrm{TXOP\ limit}$, as well as an arbitrary backoff multiplier β. The accuracy of the model was verified by comparison with ns-2 simulation. Our numerical results demonstrate that the model is accurate over a wide range of parameter settings, encompassing configurations where differentiation is provided by one or multiple differentiation mechanisms. Not only do the mean and standard deviation values match well with simulation, but the distribution values obtained by numerical inversion are in remarkably good agreement down to small tail probabilities.

Using the model, we derived 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 ratio of the burst sizes and its square root, respectively. We also used the model to study the effectiveness of CWmin, AIFS, TXOP, and β differentiation. We found 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 (except for low load) which leads to fairly predictable behavior. Differentiation based on β leads to greater dispersion in the delay of the low priority class, which is desirable.

APPENDIX

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

A. Derivation of (49)

We approximate the mean delay as follows:

$$\begin{aligned} \mathrm{E}[D] &\approx \mathrm{E}[A] & (56)\\ &\approx \sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]\, \mathrm{E}[Y], & (57) \end{aligned}$$

where (56) follows because, when c is large, backoff windows become large and more interruptions occur, and so $\mathrm{E}[A]$ dominates the other terms. Similarly, (57) 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 (57) further, we now derive approximations for $\mathrm{E}[Y]$ and $\sum_{i=0}^{R-1} \eta c^i \sum_{j=0}^{i} \mathrm{E}[U_j]$.

The $\mathrm{E}[Y]$ can be written as

$$\mathrm{E}[Y] = (1-c)t_{slot} + c\,\mathrm{E}[G] \approx c\,\mathrm{E}[G] \approx \frac{c\Gamma}{q^h}, \qquad (58)$$

where $\Gamma := \mathrm{AIFS}_1 + t_{data} + \mathrm{SIFS} + t_{ack}$, and $h := h^{(k)} = (\mathrm{AIFS_k} - \mathrm{AIFS}_1)/t_{slot}$. Equation (58) 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

$$\begin{aligned}\sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\mathrm{E}[U_j] &\approx \sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\frac{W}{2}\beta^j \qquad (59)\\ &\approx \frac{1}{\beta-1}\frac{W}{2}\beta\sum_{i=0}^{R-1}\eta c^i\beta^i, \qquad (60)\end{aligned}$$

where (59) follows under the assumption $W \gg 1$, (60) is obtained because for sufficiently large c and β, it can be shown that

$$\beta\sum_{i=0}^{R-1}\eta c^i\beta^i \gg 1. \tag{61}$$

Finally, by substituting (58) and (60) into (57), and using the approximation

$$\frac{1}{p}=\sum_{i=0}^{R-1}\eta c^i\,\mathrm{E}[U_i] \;\approx\; \frac{W}{2}\sum_{i=0}^{R-1}\eta c^i\beta^i, \tag{62}$$

we obtain (49).

B. Derivation of (50)

We approximate the variance as

$$\begin{aligned}\mathrm{V}[D] &\approx \mathrm{V}[A] \qquad (63)\\ &\approx \sum_{i=0}^{R-1}\eta c^i\{\sum_{j=0}^{i}\mathrm{V}[B_j]+\mathrm{E}[A_i]^2\}-\mathrm{E}[A]^2 \qquad (64)\\ &\approx \sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\mathrm{E}[U_j]\,\mathrm{V}[Y]+\sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\mathrm{V}[U_j]\,\mathrm{E}[Y]^2\\ &\quad+\sum_{i=0}^{R-1}\eta c^i(\sum_{j=0}^{i}\mathrm{E}[U_j]\,\mathrm{E}[Y])^2-\mathrm{E}[A]^2, \qquad (65)\end{aligned}$$

where (63) follows because $\mathrm{V}[A] \gg \mathrm{V}[\epsilon]$ (since c is large), (64) 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

$$\begin{aligned}\mathrm{V}[Y] &\approx c(1-c)\,\mathrm{E}[G]^2+c\,\mathrm{V}[\epsilon] \qquad (66)\\ &\approx c(1-c)\,\mathrm{E}[G]^2+(1-q^h)\,\mathrm{E}[G]^2\\ &\approx c(2-c)\,\mathrm{E}[G]^2, \qquad (67)\\ \mathrm{V}[U_j] &= \frac{\beta^{2j}W^2-1}{12}\approx\frac{W^2}{12}\beta^{2j}, \qquad (68)\end{aligned}$$

where, (66) is again because $\mathrm{E}[G] \gg t_{slot}$, (67) follows from $q^h \ll 1$, and (68) is due to $W \gg 1$. To further simplify (65), we note that for large R and sufficiently large β, it can be shown that

$$\begin{aligned}\sum_{i=0}^{R-1}\eta c^i\beta^{2i} &\gg \sum_{i=0}^{R-1}\eta c^i\beta^i, \qquad (69)\\ \sum_{i=0}^{R-1}\eta c^i\beta^{2i} &\gg (\sum_{i=0}^{R-1}\eta c^i\beta^i)^2. \qquad (70)\end{aligned}$$

Based on (61), (69) and (70), and under the assumption that $W \gg 1$, we obtain

$$\begin{aligned}\sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\mathrm{V}[U_j] &\approx \sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\frac{\beta^{2j}W^2}{12}\\ &\approx \frac{W^2\beta^2}{12(\beta^2-1)}\sum_{i=0}^{R-1}\eta c^i\beta^{2i}, \qquad (71)\\ \sum_{i=0}^{R-1}\eta c^i(\sum_{j=0}^{i}\mathrm{E}[U_j])^2 &\approx \sum_{i=0}^{R-1}\eta c^i(\sum_{j=0}^{i}\frac{W}{2}\beta^j)^2\\ &\approx \frac{W^2\beta^2}{4(\beta-1)^2}\sum_{i=0}^{R-1}\eta c^i\beta^{2i}. \qquad (72)\end{aligned}$$

Finally, substituting (57), (58), (60), (67), (68), (71) and (72) into (65), and using the fact of (69) and (70), we obtain

$$\begin{aligned}\mathrm{V}[D] &\approx \frac{W^2\beta^2}{12(\beta^2-1)}c^2\,\mathrm{E}[G]^2\sum_{i=0}^{R-1}\eta c^i\beta^{2i}\\ &\quad+\frac{W^2\beta^2}{4(\beta-1)^2}c^2\,\mathrm{E}[G]^2\sum_{i=0}^{R-1}\eta c^i\beta^{2i} \qquad (73)\\ &\approx \frac{c^2W^2\Gamma^2}{q^{2h}}\frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2}\sum_{i=0}^{R-1}\eta c^i\beta^{2i}. \qquad (74)\end{aligned}$$

C. Derivation of (53) and (54)

The mean delay is given by

$$\begin{aligned}\mathrm{E}[D] &\approx \frac{\mathrm{E}[D_1]}{N} \qquad (75)\\ &\approx \frac{\mathrm{E}[Y]\sum_{i=0}^{R-1}\eta c^i\sum_{j=0}^{i}\mathrm{E}[U_j]}{N}, \qquad (76)\end{aligned}$$

where (75) follows because $\mathrm{E}[D_1] \gg (N-1)(\mathrm{SIFS}+t_{data})$ when c is sufficiently large, and (76) comes from (57). We approximate $\mathrm{E}[Y]$ as follows:

$$\begin{aligned}\mathrm{E}[Y] &= (1-c)t_{slot}+\gamma_1\,\mathrm{E}[G_1]+\gamma_2\,\mathrm{E}[G_2]+\nu\,\mathrm{E}[H]\\ &\approx [c+(n_1N_1+n_2N_2-n)p(1-p)^n]\,\Gamma, \qquad (77)\end{aligned}$$

where $\Gamma := t_{data}+\mathrm{SIFS}+t_{ack}+\mathrm{AIFS}$ (note that $\mathrm{AIFS}=\mathrm{AIFS}_1=\mathrm{AIFS}_2$ for TXOP only differentiation).

Substituting (60) and (77) into (76), and making use of (62), yields (53).

The variance of the access delay can be simplified as:

$$\begin{aligned}\mathrm{V}[D] &\approx \frac{\mathrm{V}[D_1]+\frac{N-1}{N}\,\mathrm{E}[D_1]^2}{N} \qquad (78)\\ &\approx \frac{\mathrm{V}[D_1]+\mathrm{E}[D_1]^2}{N} \qquad (79)\\ &\approx \frac{\sum_{i=0}^{R-1}\eta c^i\{\sum_{j=0}^{i}\mathrm{V}[B_j]+\mathrm{E}[A_i]^2\}}{N}, \qquad (80)\\ &\approx \frac{(2\beta+1)\beta^2}{6(\beta+1)(\beta-1)^2}\frac{W^2\,\mathrm{E}[Y]^2}{N}\sum_{i=0}^{R-1}\eta c^i\beta^{2i}, (81)\end{aligned}$$

where (78) is obtained using the fact that $\mathrm{E}[D_1] \gg t_{data}+\mathrm{SIFS}$, (79) follows from $N(\mathrm{V}[D_1]+\mathrm{E}[D_1]^2) \gg \mathrm{E}[D_1]^2$ for sufficiently large β, (80) follows from (56) and (64), and (81) is obtained via similar arguments that led from (64) to (74). Finally, substituting (77) into (81) leads to (54).

References

[1] IEEE, *IEEE 802.11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, 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*, 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 Transactions on Wireless Communications*, 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 Communications and Networking Conference*, vol. 4, pp. 1938–1943, 2006.

[6] ——, "MAC access delay in IEEE 802.11e EDCA," *Proc. IEEE 64th Vehicular Technology Conference - Fall*, 2006.

[7] Y. Xiao, "Performance analysis for priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Transactions on Wireless Communications*, 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 Journal on Selected Areas in Communications*, vol. 22, no. 10, pp. 2095–2106, 2004.

[9] J. Tantra, C. H. Foh, and A. Mnaouer, "Throughput and delay analysis of the IEEE 802.11e EDCA saturation," *Proc. IEEE International Conference on Communications*, vol. 5, pp. 3350–3454, 2005.

[10] X. Tao and S. Panwar, "Throughput and delay analysis for the IEEE 802.11e enhanced distributed channel access," *IEEE Transactions on Communications*, vol. 54, no. 4, pp. 596–603, 2006.

[11] T.-C. Tsai and M.-J. Wu, "An analytical model for IEEE 802.11e EDCA," *Proc. IEEE International Conference on Communications*, vol. 5, pp. 3474–3478, 2005.

[12] J. Hui and M. Devetsikiotis, "A unified model for the performance analysis of IEEE 802.11e EDCA," *IEEE Transactions on Communications*, vol. 53, no. 9, pp. 1498–1510, 2005.

[13] H. Zhu and I. Chlamtac, "Performance analysis for IEEE 802.11e EDCF service differentiation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1779–1788, 2005.

[14] P. E. Engelstad and O. N. Østerbø, "Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction," *Proc. the 8th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems (ACM MSWiM)*, pp. 224–233, 2005.

[15] ——, "Analysis of the total delay of IEEE 802.11e EDCA and 802.11 DCF," *IEEE International Conference on Communications*, vol. 2, pp. 552–559, 2006.

[16] J.-D. Kim and C.-K. Kim, "Performance analysis and evaluation of IEEE 802.11e EDCF," *Wireless Communications and Mobile Computing*, vol. 4, no. 1, pp. 55–74, 2004.

[17] J. W. Robinson and T. S. Randhawa, "Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 917–928, 2004.

[18] ——, "A practical model for transmission delay of IEEE 802.11e enhanced distributed channel access," *Proc. IEEE International Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 1, pp. 323–328, 2004.

[19] F. Peng, H. M. Alnuweiri, and V. C. M. Leung, "Analysis of burst transmission in IEEE 802.11e wireless LANs," *Proc. IEEE International Conference on Communications*, vol. 2, pp. 535–539, 2005.

[20] 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.

[21] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[22] V. Ramaiyan, A. Kumar, and E. Altman, "Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness, multistability and throughput differentiation," *Proc. SIGMETRICS'05*, 2005.

[23] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke, "Analysis of IEEE 802.11e for QoS support in wireless LANs," *IEEE Wireless Communications*, vol. 10, no. 6, pp. 40–50, Dec. 2003.

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

[25] 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.

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

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