Report No. UIUCDCS-R-2006-2711 **UILU-ENG-2006-1743**

A Comprehensive Study of the IEEE 802.11e Enhanced Distributed Control Access (EDCA) Function

by

Chunyu Hu and Jennifer C. Hou

April 2006

A Comprehensive Study of the IEEE 802.11e Enhanced Distributed Control Access (EDCA) Function

Chunyu Hu† and Jennifer C. Hou‡
† Department of Electrical and Computer Engineering
‡ Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA
E-mail:{*chunyuhu, jhou*}@*uiuc.edu*

Abstract

This technical report presents a comprehensive study of the Enhanced Distributed Control Access (EDCA) function defined in IEEE 802.11e. All the three factors are considered. They are: *contention window size (CW)*, *arbitration inter-frame space (AIFS)*, and *transmission opportunity limit (TXOP)*. We first propose a discrete Markov chain model to describe the channel activities governed by EDCA. Then we evaluate the individual as well as joint effects of each factor on the throughput and QoS performance. We obtain several insightful observations showing that judiciously using the EDCA service differentiation mechanism is important to achieve maximum bandwidth utilization and user-specified QoS performance. Guided by our theoretical study, we devise a general QoS framework that provides QoS in an optimal way. The means of realizing the framework in a specific network is yet to be studied.

I. Introduction

Recent years has observed the proliferation of computers and a large variety of consumer electronics devices such as PDAs, cell phones, digital camera/recorders and digital home media centers. The later raises a high demand of ubiquitous inter-connecting among these devices and/or the Internet. The rapid advance and maturity of wireless technology makes it possible to meet this ever-increasing and ever-evoluting demand. The success of 802.11-compliant wireless networks is an evidence. The diversity of applications and their co-existence impose a stringent requirement for Quality of Service (QoS). A number of factors contribute to this: first, data transfer and real-time video/audio streaming have different traffic characteristics. The former is bursty in nature and the later is periodic and delay sensitive. Second, interactive applications such as SSH, web browsing are more sensitive to delay than bulk data transfer such as ftp does. Third, real-time systems, such as distributed control system, have strict deadline requirement. Finally, users may different preferences and/or pay different prices and they expect services correspondingly delivered. As the result, wireless networks have to be designed to support applications with service differentiation. Foreseeing the need, the IEEE 802.11 Task Group E has approved the standard IEEE 802.11e, as the QoS enhancements of the IEEE 802.11 MAC protocol.

The 802.11 MAC protocol defines two access methods: Distributed Coordination Function (DCF) and Point Coordination Function (PCF). DCF is fully distributed. It assumes a binary exponential backoff algorithm to resolve contention and collisions. PCF is poll-based: the central coordinator (which is usually an access point) polls each station to give it the right to transmit a packet. 802.11e extends DCF and PCF with Enhanced Distributed Channel Access (EDCA) and HCF (Hybrid Coordination Function) Controlled Channel Access (HCCA), respectively. Both EDCA and HCCA defines Traffic Classes and each class is assigned of a priority. A detailed overview on 802.11e will be provided in Section II. Because distributed access involves little or no central coordination, it is more widely used. Therefore, we will focus on EDCA in our study.

Technical Report No. UIUCDCS-R-2006-2711 (Engr. No. UILU-ENG-2006-1743), April 2006

In EDCA, medium access is contention-based using the same backoff algorithm as DCF and is prioritized by three configurable parameters: the contention window size (*CW*), the arbitration inter frame space (*AIFS*) and the transmission opportunity limit (*TXOP*). CW and AIFS determine the probability of gaining the channel access, while TXOP determines the time of occupying the channel *after* the channel access is obtained. To explain the former, every time a backoff procedure is initiated, the backoff time (in number of slots) is uniformly generated in $[0, CW-1]$. A station has to backoff this amount of time before a transmission attempt is made. AIFS defines the amount of time that has to be sensed idle before the backoff procedure is initialized/resumed. See Fig. 2 for an illustration. Flows of different priorities are assigned different parameter values to increase/decrease their chance of gaining medium access. Although this is intuitively correct, it is important to understand quantitatively how, and to what extent, the two parameters favor/disfavor data transmission from high-priority/low-priority flows. In this report, we first study the effect of CW and AIFS, and then discuss the effect of TXOP.

Several studies on evaluating the performance of EDCA have been made *via simulation* in the context of IEEE 802.11e in [4], [10], [11], [12] and [14]. Several theoretical models that shed insights on how service differentiation can be achieved have been reported in [5], [8], [9], [15], [16], [19], and [20], again in the context of IEEE 802.11e. Most of the models, if not all, are based on Bianchi's model [2] or Calí's model [3] that were proposed to study the performance of IEEE 802.11 DCF under the asymptotic condition (i.e., all the stations always have packets ready for transmission). Among all the models, those reported in [9] and [19] analyze the effect of varying the contention window size on the performance of service differentiation, and those reported in [5], [8], [15], [16], and [20] also study the effect of varying AIFS values on the performance. In most (if not all) of the work that study the AIFS effect, such as [8], [15], [16], and [20] a different contention window range $[1, CW]$ (rather than $[0, CW-1]$) is assumed. As has been pointed out in [6], this subtle difference results in considerable degradation in the system throughput. In summary, a model that conforms to the standard and fully incorporates both parameters is yet to observe. A joint study based on that on how, and to what extend, the two parameters affect the performance is expected.

In this report, we propose a comprehensive, accurate, and yet elegant model to characterize data activities in EDCA and jointly study the effects of varying CW and AIFS. Based on our analytical model of IEEE 802.11 DCF, instead of using a p-persistent model (that is commonly assumed in literature), we use a discrete-time Markov chain to describe the transition of the channel state. We validate our analytical model and evaluate the individual as well as joint effects of each factor (CW, AIFS and TXOP) via simulation and theoretical analysis. We obtain several insightful observations from the above study, namely, 1) Traffic classes with small values of AIFS dominate channel access, depriving traffic classes with larger AIFS values of their channel access. 2) With the currently proposed parameter setting, the differentiation mechanism fails to allocate bandwidth among stations of different classes at deterministic QoS. That is, given the currently proposed parameter setting, the default EDCA parameter settings can not fully support both real-time data streams and best-effort applications at their configured QoS. And 3) The available bandwidth is under-utilized given the proposed parameter setting. These observations enlighten us to devise a general QoS framework to provide QoS in an optimal way in the sense that it can achieve maximal bandwidth utilization as well as user-specified QoS performance (for both delay-sensitive real-time traffic and prioritized best-effort traffic).

The rest of the report is organized as follows. Section II gives an overview of IEEE 802.11e EDCA. In Section III, following an introduction of the network model and the assumptions, we present the analytical model that characterizes the EDCA service differentiation mechanism. In Section IV, we carry out simulation to validate the model as well as to evaluate the performance of EDCA. In Section VI we devise a general QoS provisioning framework that can achieve maximal bandwidth utilization and satisfactory QoS performance. Finally, we conclude the report in Section VII.

II. An Overview of 802.11e EDCA

IEEE 802.11 defines the basic contention-based access method, DCF; however, it does not provide any differentiated service. This limits its application to situations, for example, where real-time multimedia traffic requires delay-constraint service or more generally, traffic flows generated by various applications have different priorities.

To deal with this inadequacy, the IEEE 802.11 Working Group E was formed and makes an extension to the legacy IEEE 802.11 MAC, IEEE 802.11e.

In IEEE 802.11e, in addition to DCF and PCF, a new function, Hybrid Coordination Function (HCF) is introduced. It extends DCF and PCF, respectively, with an Enhanced Distributed Channel Access (EDCA) and a Controlled Channel Access (HCCA). The new MAC architecture is depicted in Fig. 1.

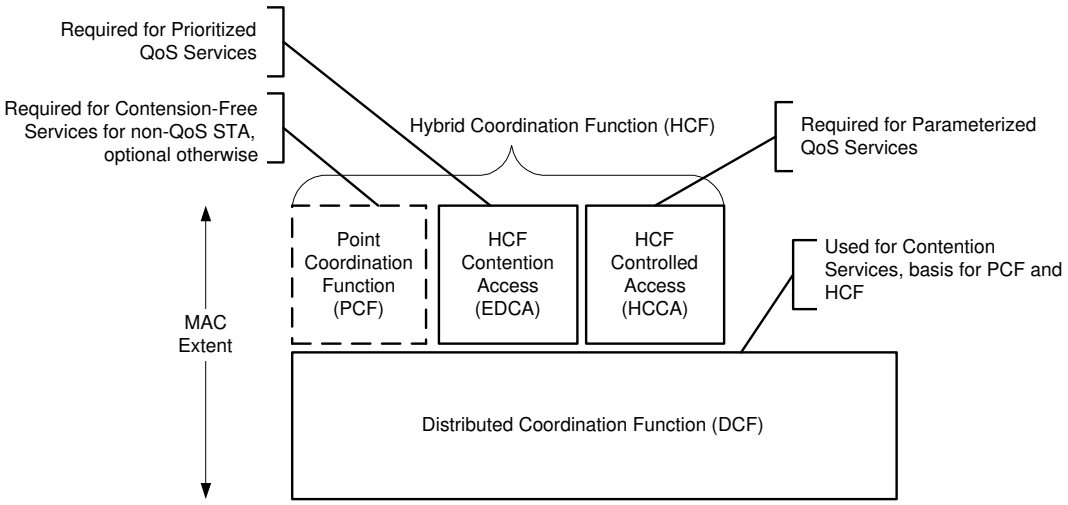


Fig. 1. IEEE 802.11e MAC architecture.

EDCA is the QoS enhancement of DCF. In EDCA, the traffic is categorized according to its priority. All the stations access the channel using the same backoff algorithm as in DCF, but are configured with a set of QoS parameters associated with each priority. They are: contention window size (CW), arbitration inter frame space (AIFS, in replacement of DIFS), and transmission opportunity (TXOP) limit. Similarly to DCF, CW determines the number of backoff slots (which is uniformly distributed in [0, CW-1]) a station has to count down before a transmission attempt is made. AIFS determines the duration that has to be sensed idle before the backoff procedure is initialized/resumed. It is specified in a way similar to DIFS as follows:

$$\text{AIFS[AC]} = \text{SIFS} + \text{AIFSN[AC]} \times \text{aSlotTime}, \tag{1}$$

where AIFSN[AC] is the arbitration inter frame space number. Fig. 2 illustrates the relation between some inter frame spaces. Therefore, CW and AIFS determine the chance of obtaining the channel access. Generally, the higher

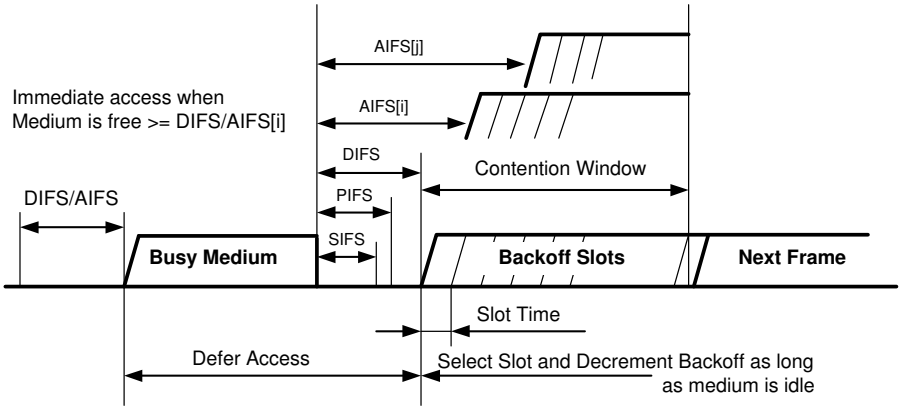


Fig. 2. The relation between some inter frame spaces.

priority a class has, the smaller its CW and/or AIFS values. On the other hand, the TXOP limit enables the block acknowledgment following a normal successful DATA-ACK transmission. It determines the time of occupying the channel after the access is obtained. The default EDCA parameter setting defines eight priority classes as shown in Tab. I, which are mapped into four access categories and their parameters are specified in Tab. II.

TABLE I

802.11E: USER PRIORITY TO ACCESS CATEGORY MAPPINGS.

Priority	User Priority	802.1D Designation	Access Category	Designation
lowest	1	BK	AC_BK	Background
	2	-	AC_BK	Background
	0	BE	AC_BE	Best Effort
	3	EE	AC_BE	Video
	4	CL	AC_VI	Video
	5	VI	AC_VI	Video
	6	VO	AC_VO	Voice
highest	7	NC	AC_VO	Voice

TABLE II

802.11E: DEFAULT EDCA PARAMETER SETTINGS.

AC	CWmin	CWmax	AIFSN	TXOP Limit		
				DS_CCK	Extended Rate/OFDM	Other PHYS
AC_BK	32	1024	7	0	0	0
AC_BE	32	1024	3	0	0	0
AC_VI	16	32	2	6.016ms	3.008ms	0
AC_VO	8	16	2	3.264ms	1.504ms	0

III. AN ANALYTICAL MODEL FOR EDCA

A. Network Model and Assumptions

We consider a general single-cell wireless network without any central coordinator. All stations can hear each other, i.e., there exists no hidden terminal. It is assumed that the channel is in the ideal condition, meaning that it does not introduce any errors other than those caused by collisions. No capture effect is considered. We assume all stations are back-logged, having packets to transmit. It is called *the asymptotic condition*, and is a common assumption made for theoretical tractability to analyze the saturation performance.

There are M priority classes, with the number of stations in each class being N_j, $j = 1, \cdots, M$. Each class is configured with a set of QoS parameters for distributed access contention: the inter-frame space $AIFS_j$ and the contention window size CW_j (the parameter TXOP limit will be considered at a later time). Without loss of generality, we assume $AIFSN_1 \leq AIFSN_2 \leq \cdots \leq AIFSN_M$. In particular, let m denote the priority index such that $AIFSN_1 = \cdots = AIFSN_m < AIFSN_{m+1} \leq \cdots \leq AIFSN_M$.

All stations share and access the channel with the use of binary exponential backoff algorithm. That is, a random integer value is uniformly chosen in $[0, CW_i - 1]$ and used to initialize the backoff timer, where CW_i is the current contention window for traffic class i. The backoff timer is decremented as long as the channel is sensed idle, frozen when data transmission (initiated by other stations) is in progress, and resumed when the channel is sensed idle again for more than $AIFS_i$, where i denotes the traffic class. The time immediately following an idle period of length *short inter frame space* (SIFS) is slotted, with each slot equal to the time needed for any station to detect the transmission of a packet from any other station. When the backoff timer expires, the station attempts for frame transmission at the beginning of the next slot. Finally, if the data frame is successfully received, the receiver transmits an acknowledgment frame after a SIFS, On the other hand, if an acknowledgment has not been received when the

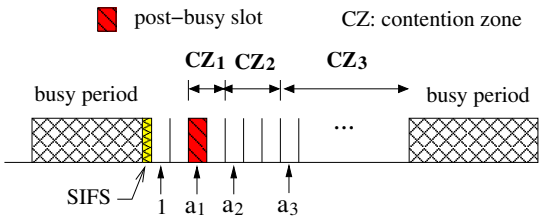


Fig. 3. An illustration of the notions of contention zones and the post-busy slot. In this example, there are three classes with different AIFS values. $a_j = AIFSN_j + 1$.

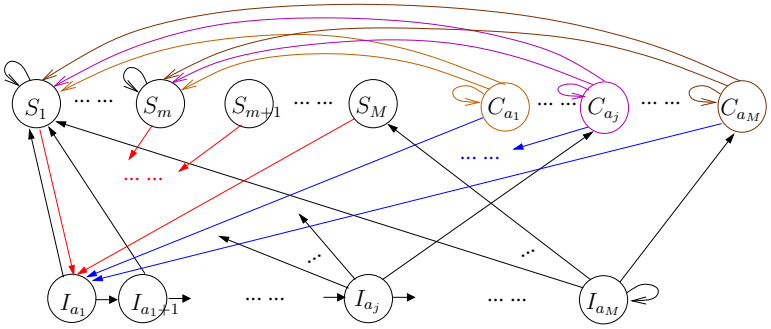


Fig. 4. The discrete Markov chain that describes the channel state transition

sender times out, the data frame is presumed to be lost, and a retransmission is scheduled. The value of CW_i is set to $CW_{min}(i)$ in the first transmission attempt, and is doubled at each retransmission up to value $CW_{max}(i)$. A data frame of class i is retried for a maximum number of times, denoted by L_i, upon transmission failure.

Let $a_j \stackrel{\triangle}{=} AIFSN_j + 1$. For ease of exposition, we divide the slots subsequent to a busy period into consecutive *contention zones*. Specifically, as illustrated in Fig. 3, *contention zone* j ($j = 1, \cdots, M-1$) starts at the a_j-th slot and ends at (including) the $(a_{j+1}-1)$-th slot. The *M-th contention zone* includes the a_M-th slot and all the slots beyond. An important observation that will be used throughout the derivation is — under EDCA, *only stations of the first j classes are eligible to transmit in the j-th contention zone.* Also, as the a_1-th slot is the first slot immediately following a busy period and (as will be shown later) plays an important role in the performance, we term it as the *post-busy slot*.

B. Channel State Space

For ease of explanation, we treat a collision period or a successful transmission as a virtual (busy) slot. The channel state, denoted by $s(t)$, is sampled at the end of each busy/idle slot. There are three types of possible channel states:

1) State I_k ($k = a_1, a_1+1, \cdots, a_M-1$): the idle channel state in the k-th slots after the busy slot. (Recall that the first slot is the one immediately after a busy slot plus a SIFS.) In other words, when the channel is in I_k, it means there are $k-1$ consecutive idle slots subsequent to a busy slot and the k-th slot is still idle. In addition, I_{a_M} is the channel state in which there are $\geq a_M$ consecutive idle slots.
2) State S_j: a successful transmission made by a station of class j.
3) State C_{a_j}: either a collision subsequent to an idle slot that is in the j-th contention zone or a collision subsequent to such a collision. By the definition of the contention zone, collision C_{a_j} only involves stations of the first j classes.

The channel state space is thus defined as $\mathbb{S} = \{ I_k,\ S_j,\ C_{a_j} : j = 1, \cdots, M \text{ and } a_1 \leq k \leq a_M \}$.

C. Transitions of Channel States

We use a discrete-time Markov chain (Fig. 4) to describe the transition of channel states. Possible transitions are:

1) $I_k \rightarrow \{\ I_{k+1}, S_1, \cdots, S_j, C_{a_j}\ \}$, for $a_j \leq k \leq a_{j+1}-1$, $j = 1, \cdots, M-1$: An idle slot state can transition to another idle slot state, a collision state, or a successful transmission state. Specifically, an idle slot state in the j-th contention zone can transition to a successful transmission (made by one of the stations of the first j classes), or to a collision C_{a_j} (caused by two or more stations of the first j classes attempting for transmission), or to the next idle slot in the time order.
2) $I_{a_M} \rightarrow \{\ I_{a_M}, S_1, \cdots, S_M, C_{a_M}\ \}$: The only exception in the transition from an idle slot state occurs when the idle slot is in the M-th contention zone. By the definition of the M-th contention zone, an idle slot state in the M-th contention zone can transition to itself. Hence, as shown in Figure 4, instead of having an outgoing arrow into other idle slot states, state I_{a_M} contains a self-loop pointing to itself.
3) $S_j \rightarrow \{\ S_j, I_{a_1}\ \}$, for $j = 1, \cdots, m$: As a station will transmit immediately once its backoff timer counts down to zero, when the backoff timer freezes upon detection of the busy medium, the timer value is *always* positive and has at least 1 slot time. Therefore, stations that did not participate in transmission in the busy period will *not* transmit in the post-busy slot (i.e., the a_1-th slot immediately after a busy period). This peculiar access behavior to the *post-busy* slot is often ignored in previous work, which assumes uniform and independent access to *any slot* (*a.k.a.* p-persistent). Because of this behavior, subsequent to a successful transmission, either another successful transmission made by the *same* station follows, or the channel becomes idle. This implies, for $j \leq m$, state S_j can either transition to itself or the idle slot state I_{a_1}.
4) $S_j \rightarrow \{\ I_{a_1}\ \}$, for $j = m+1, \cdots, M$: As stations of classes $j > m$ are assigned larger AIFS values, they are not eligible to access the *post-busy* slot. Hence, states S_j, $j > m$, will transit exclusively to the idle slot state I_{a_1}.
5) $C_{a_j} \rightarrow \{\ C_{a_j}, I_{a_1}, S_1, \cdots, S_m\ \}$, for $j = 1, \cdots, M$: Transitions from collision states can be similarly explained as in cases 3–4 and hence are not elaborated on here.

D. Derivation of State Transition Probabilities

Now we are in a position to derive the transition probabilities of all the possible transitions. We assume that stations of class j access a slot other than an *post-busy* slot *independently* and *uniformly* with probability τ_j. τ_j is termed as the *attempt probability*. For ease of exposition, we first consider the case that the contention window size CW_j is fixed. Then we extend the model to accommodate that case that the contention window size changes in compliance with the the binary exponential backoff procedure.

Before we proceed, we define the following terms for notational convenience:

$$A_j \triangleq \prod_{h=1}^{j}(1-\tau_h)^{N_h}, \quad B_j \triangleq \prod_{h=1}^{j}\left(1-\frac{\tau_h}{CW_h}\right)^{N_h}. \tag{2}$$

a) Transitions from the idle slot state to other states: Recall that in the j-th contention zone, only the first j classes are eligible to contend for channel access. Hence, we derive, for each contention zone, transition probabilities from an idle channel state. For notational convenience, we define $a_{M+1} = a_M + 1$.

For $k \in [a_j, a_{j+1})$, $j = 1, 2, \cdots, M$, and $u \leq j$,

$$P[I_k \rightarrow I_{k+1}] = A_j, \tag{3}$$

$$P[I_k \rightarrow S_u] = N_u\tau_u(1-\tau_u)^{N_u-1}\prod_{h=1,h\neq u}^{j}(1-\tau_h)^{N_h} = \frac{N_u\tau_u}{1-\tau_u}A_j \tag{4}$$

$$P[I_k \rightarrow C_{a_j}] = 1 - P[I_k \rightarrow I_{k+1}] - \sum_{u=1}^{j}P[I_k \rightarrow S_u] \tag{5}$$

b) Transitions from a successful transmission state to other states: After a station of the first m classes finishes a successful transmission, it may gain the channel access again if it chooses 0 as the next backoff timer value. This occurs with probability $\frac{1}{CW_j}$, since the backoff timer value is selected uniformly in $[0, CW_j - 1]$. On the other hand, if the station is of of class $m+1$ to M, it is not eligible to access the *post-busy* slot. Moreover, all

the other stations have frozen their backoff timer with the remaining timer value at least 1 slot, and hence will not attempt to transmit either. In this case, the *post-busy* slot is idle with probability 1. We have

$$P[S_j \to S_j] = \frac{1}{CW_j} \text{ , for } j = 1, \cdots, m, \tag{6}$$

$$P[S_j \to I_{a_1}] = \begin{cases} 1 - \frac{1}{CW_j} \text{ , for } j = 1, \cdots, m, \\ 1 \text{ , for } j = m+1, \cdots, M. \end{cases} \tag{7}$$

c) Transitions from a collision state to other states: Recall that by the definition of m, $AIFSN_1 = \cdots = AIFSN_m$, i.e., the first m contention zones is practically the same one. Consequently, we merge the collision states C_{a_k}, $k = 1, \cdots, m$, into one state, and denote it by C_{a_m}. The transition probabilities originating from C_{a_j}, for $j = m, \cdots, M$ and $u \leq m$ are given below, with their detailed derivation given in Lemma 1–2 in Appendix VIII.

For $j = m, \cdots, M$ and $u \leq m$,

$$P[C_{a_j} \to I_{a_1}] = \frac{1}{P[I_k \to C_{a_j}]} \left\{ B_m - A_j \left[1 + \sum_{k=1}^{m} \frac{N_k \tau_k}{1 - \tau_k} \times \left(1 - \frac{1}{CW_k} \right) + \sum_{k=m+1}^{j} \frac{N_k \tau_k}{1 - \tau_k} \right] \right\}, \tag{8}$$

$$P[C_{a_j} \to S_u] = \frac{1}{P[I_k \to C_{a_j}]} N_u \frac{\tau_u}{CW_u} \left(\frac{B_m}{1 - \frac{\tau_u}{CW_u}} - \frac{A_j}{\tau_u} \right), \tag{9}$$

$$P[C_{a_j} \to C_{a_j}] = 1 - P[C_{a_j} \to I_{a_1}] - \sum_{u=1}^{m} P[C_{a_j} \to S_u]. \tag{10}$$

With all the derived transition probabilities, we can compute the stationary probabilities of channel states by solving the equilibrium equations for the Markov chain, $\mathbf{s} = \mathbf{s}\mathbb{P}$. Let the equilibrium channel state be denoted by $\widetilde{s}$.

E. Derivation of the Attempt Probability

Recall that each station attempts to transmit in a slot (other than the *post-busy* slot) independently and uniformly with probability τ_j, where j is the priority class which the station belongs to.

Given a fixed contention window size, CW_j, for each class j, τ_j can be simply expressed as $\frac{2}{CW_j}$ – the inverse of the average waiting (backoff) time. Note that the backoff timer is frozen when data transmission (initiated by other stations) is in progress, and resumed when the channel is sensed idle again for more than AIFS[i]. The expression results from that the backoff time at the beginning of any *eligible* slot is approximately uniformly distributed in $[0, CW_j - 2]$, where an *eligible* slot refers to any non-*post-busy* slot.

An iterative algorithm to derive $\overline{CW_j}$ and τ_j: In the case that the contention window size CW_j changes in compliance with the binary exponential backoff procedure, we develop an iterative algorithm to derive the average contention window size $\overline{CW_j}$.

Consider the view of a tagged station of class j. Let $p_{coll}(j)$ denote the probability that when the tagged station transmits a frame in an eligible slot, the frame incurs a collision; and $p_{coll}^{(r)}(j)$ denotes this probability in the r-th iteration. The average contention window size, $\overline{CW_j}^{(r)}$, in the r-th iteration can be computed from its probability mass function:

$$\begin{aligned} q(j, \ell) &\stackrel{\triangle}{=} P\left[CW_j^{(r)} = W(j, \ell) \right] \\ &= p_0 \left(p_{coll}^{(r)}(j) \right)^{\ell} \text{ for } \ell = 0, \ldots, L_j, \end{aligned} \tag{11}$$

where p_0 is the normalization factor, and can be obtained by noting $\sum_{\ell=0}^{L_j} q(j, \ell) = 1$. L_j is the retry limit for class j and $W(j, \ell) = \min\{2^{\ell} CW_{min}(j), CW_{max}(j)\}$, $\ell = 0, \cdots, L_j$. The attempt probability for the next iteration, $\tau_j^{(r+1)}$, can be computed from $\overline{CW_j}^{(r)}$ by $\tau_j^{(r+1)} = \frac{2}{\overline{CW_j}^{(r)}}$.

The probability $p_{coll}(j)$ is yet to be derived. As this probability varies in different contention zones, we will first derive the conditional probability of collision given that the system is in the contention zone k ($j \leq k \leq M$),

and then the probability that the system is in the contention zone k. The former probability can be expressed as $p_{coll}(j,k) = 1 - \frac{A_k}{1-\tau_j}$, i.e., the probability that at least one other station of the first k classes transmits in the same slot. The latter probability that the system is in the contention zone k is $\sum_{h=a_k}^{a_{k+1}-1} \mathbb{P}[\widetilde{s} = I_h]$. Now $p_{coll}(j)$ can be expressed as

$$p_{coll}(j) = \frac{1}{c_0} \sum_{k=j}^{M+1} \left(1 - \frac{A_k}{1-\tau_j}\right) \left(\sum_{h=a_k}^{a_{k+1}-1} \mathbb{P}[\widetilde{s} = I_h]\right), \tag{12}$$

where $c_0 = \sum_{h=a_k}^{a_M} \mathbb{P}[\widetilde{s} = I_h]$ and recall $a_{M+1} \stackrel{\Delta}{=} a_M + 1$. Note that all the stationary probabilities used here should be derived from the perspective of of the tagged station, i.e., the number of stations in the class which the tagged station belongs to is reduced by 1 in all the relevant calculation.

The average attempt probabilities calculated in the iterative algorithm will replace all the τ_j terms in Eqs. (6)-(10). Since after a successful transmission, a station will reset its contention window size to the minimum value, CW_j in Eqs. (3)-(10) will be replaced by $CW_{min}(j)$ after a successful transmission.

F. Derivation of the System Throughput

We compute the system throughput by calculating the average amount of successful transmission (in bits) over the expected length of a slot (By a *slot*, we mean either a successful transmission, a collision, or an idle slot). Specifically, let t_s denote the length of an idle slot (which is a PHY parameter), and T_D and T_C, respectively, the average length of a successful transmission and the average length of a collision period. With the results derived in Sections III-C and III-E, the expected slot time can be expressed as

$$\overline{t} = t_s \sum_{k=a_1}^{a_M} \mathbb{P}[\widetilde{s} = I_k] + T_D \sum_{j=1}^{M} \mathbb{P}[\widetilde{s} = S_j] + T_C \sum_{j=1}^{M} \mathbb{P}[\widetilde{s} = C_{a_j}]. \tag{13}$$

In the basic distributed access mechanism, i.e.,, without the RTS-CTS floor acquisition mechanism, a successful transmission contains transmission of a DATA frame and a SIFS followed by an ACK. The second term results from that after each successful transmission, the backoff timer of a station is resumed only after an idle period of AIFS, and for ease of computation, we consider $AIFS_{min}$ ($\stackrel{\Delta}{=} \min\{AIFSN_j,\ j = 1, \cdots, M\}$) as part of a successful transmission, i.e., $T_D = DATA + SIFS + ACK + AIFS_{min}$, where $DATA$ is the transmission time of a data frame.

A collision is detected by a sender station upon the timeout of the sender timer, and by other stations when they receive corrupted packets. After detecting a collision, a receiver node resumes its backoff timer after an idle period of EIFS, where EIFS is set to $EIFS = SIFS + ACK + AIFS_{min}$, so that both colliding and non-colliding stations resume their backoff timers or start to sense the channel at approximately the same time. This gives $T_C = DATA_{max} + SIFS + ACK + AIFS_{min}$, where $DATA_{max}$ is the largest DATA frame incurred in the collision. In the case that the RTS-CTS mechanism is used, T_D and T_C can be derived in a similar manner.

The throughput of stations of class η_j ($j = 1, \cdots, M$) can be expressed as

$$\eta_j = \frac{\overline{m} \times \mathbb{P}[\widetilde{s} = S_j]}{\overline{t}}, \tag{14}$$

where $\overline{m}$ is the average payload (in bits) carried in a DATA frame. Note that the underlying assumption in Eq. (14) is that the size of data frames of all classes has the same distribution. It is, however, straightforward to extend Eq. (14) to accommodate the case that the size of data frames of different classes has its individual distribution.

IV. Model Validation and Performance Evaluation

We have performed a simulation study to both validate the analytic model derived in Section III and to evaluate the performance of EDCA (in conjunction with the current parameter setting as suggested in MBOA MAC [13], [17]).

The PHY and MAC parameters, as suggested in MBOA PHY/MAC proposals [13], [1] for UWB-operated WPANs, have been used in our simulation. They are listed in Table III. Note that the UWB PHY will support a rate set of {53.3, 80, 110, 160, 200, 320, 400, and 480 Mbps}, among which support for transmitting and receiving at data rates of {53.3, 110 and 200 Mbps} is mandatory. We choose the highest mandatory rate, 200 Mbps, as the data rate. Both analytical and simulation results for other data rates exhibit similar trends and thus are not reported. Each simulation run lasts for 200 simulation seconds.

TABLE III
PHY AND MAC PARAMETERS AS SUGGESTED IN MBOA UWB PHY/MAC.

Channel Rate	200 Mb/s
Basic Rate	55 Mb/s
Slot Time	8 μsec
SIFS	10 μsec
ACK Time	13.125 μsec
MPDU[1] + FCS[2]	41.25 μsec
Data Payload	1024 Bytes

[1]MPDU: Message Data Protocol Unit.
[2]FCS: Frame Check Sequence

A. Individual effect of CW

In the first set of results, we use the same AIFS value for all classes and assign different values of CW for each class. We consider three cases: In the first two cases (Fig. 5(a) and (b)), there are two priority classes each assigned a CW value (a) $CW_1 = 8, CW_2 = 16$, and (b) $CW_1 = 16, CW_2 = 32$. In the third case (Fig. 5(c)), there are 3 classes, each assigned a CW range: $CW_1 \in [8, 16]$, $CW_2 \in [16, 32]$ and $CW_2 \in [32, 64]$. Several observations are in order.

First, the analytical results agree with the simulation results very well, being in the interval $[0.9774, 1.0794]$ of the simulation results.

Second, stations that use smaller values of CW (e.g., 8 vs. 16 and 32, 16 vs. 32) grasp a large portion of available bandwidth. Clearly, a scheme that varies the values of CW is effective in allocating bandwidth in a QoS-controlled manner among different classes. (We will elaborate on how to determine the optimal values of CW to achieve deterministic proportional services differentiation in Section VI.)

Third, in general the throughput attained by each class decreases as the number of stations increases. This is consistent with our intuition, since the larger the number of stations the more likely collisions will occur. However, as shown in Fig. 5 (a), (b) and (c), the throughput curves corresponding to class 1 exhibit a peculiar trend. Specifically, in Fig. 5 (a), instead of a monotone decrease (as the number of stations in each class, N increases), the throughput increases (though slowly) between $N = 10$ and $N = 30$, before it concedes to the decreasing trend. Similar trends are observed in Fig. 5 (b) and (c), although when N is larger (the curves are cut off before the decreasing trend shows).

The above phenomenon is a result of the peculiar access behavior of the *post-busy* slot. Recall in the *post-busy* slot, only a subset of stations are eligible to contend for channel access, with a probability inversely proportional to their respective CW. Therefore we observe two access patterns: one is in the *post-busy* slot, and the other is in all the other slots. The later affects the former yet the former is independent of the later. When the collision is sparse, the probability that the *post-busy* slot is accessed is also slim. However, when collisions occur more frequently, the *post-busy* slot is also more likely to be accessed. Moreover, the probability that access to the *post-busy* slot is successful exhibits a similar trend as that for non- post-busy slots, but only lags in *phase*. Indeed we observe that when the number, N, of stations grows large, almost all non-*post-busy* slots incur collisions, and yet it is possible for a successful transmission or an idle period to occur in the *post-busy* slot. The combination of the two trends induces the interesting, peculiar fluctuation in the throughput. Note that as stations of class 1 usually have the smallest

contention windows, they dominate in accessing the post-busy slot. This explains why the peculiar trend is only observed in the aggregated throughput attained by class-1 stations.

To verify whether or not EDCA can provide deterministic (in the statistical sense) proportional QoS, we calculate, with the results shown in Fig. 5(c-1), the ratio of the throughput attained by a class $i+1$ station to that by a class i station, $i = 1, 2$. As shown in Fig. 5(c-2), instead of being managed at a stable level, the throughput ratios first decrease as N_i grows and then levels off as N_i continues to grow. Although the throughput ratio indicates that stations of higher-priority classes do attain more throughput with the use of smaller CW ranges, the throughput ratio also depends on the number of stations in the system (which usually cannot be known *a priori*). We will further discuss this issue in Section VI.

B. Individual effect of AIFS

Next we study the impact of varying AIFS values on the performance of service differentiation. We consider two priority classes, both configured with the same congestion window size $CW = 16$ but different AIFS values. The high-priority class has $AIFSN_1 = 2$, and the other has $AIFSN_2 = 3, 5$, or 7. Fig. 6 gives both the simulation results and the analytical results. Several observations are in order.

First, the analytical results agree very well with the simulation results. Second (and more importantly), stations of the high-priority class (and with smaller AIFS values) almost grasp all the available bandwidth. In particular, when the number of stations in both classes reaches 12, 6 and 4 in three cases (a), (b) and (c), respectively, the throughput attained by class-2 stations is less than 1% of the bandwidth (200 Mbps). This results from the fact that stations of class 1 can make access attempts in both contention zones 1 and 2, while stations of class 2 can only make attempts contention zone 2. This suggests that QoS provisioning by assigning different AIFS values to different access categories may lead to starvation of stations of low-priority access categories.

To explore a quantitative explanation of the phenomenon that lower-prioritized traffic gets starved, we analyze a special case of our model proposed in Section III, derive the throughput difference as a function of the difference in AIFS values. More specifically, we consider two classes with the same CW, and the same number of nodes in each class. That is, $CW_1 = CW_2 = CW$ and $N_1 = N_2 = N$. Their attempt probabilities are therefore the same as well, $\tau_1 = \tau_2 = \tau = \frac{2}{CW}$. Let $d = AIFSN_2 - AIFSN_1$ slots. We simplify some notations as follows: we use I_0 to I_{d-1} to represent the idle states in contention zone 1. States C_1 and C_2 represent respectively the collision states transited from an idle state in contention zone 1 and zone 2. The discrete Markov chain model shown in Fig. 4 is then reduced to Fig. 7. In this particular case, denote the probabilities of the equilibrium states by $\Pi = [\pi_{I_0}, \pi_{I_1}, \cdots, \pi_{I_d}, \pi_{s_1}, \pi_{s_2}, \pi_{c_1}, \pi_{c_2}]$. Dropping the I's in the subscriptions, the notation is simplified as $\Pi = [\pi_0, \pi_1, \cdots, \pi_d, \pi_{s_1}, \pi_{s_2}, \pi_{c_1}, \pi_{c_2}]$. We are interested in $\Delta = \pi_{s_1} - \pi_{s_2}$, and want to express it as a function of the variable d.

From the equilibrium equation $\Pi = \Pi P$, we have:

$$\pi_k = A_1^k \pi_0, \tag{15}$$

$$\pi_0 = (\pi_{S_1} + \pi_{S_2}) \left(1 - \frac{1}{CW}\right) + \pi_{C_1} P_{C_1,0} + \pi_{C_2} P_{C_2,0}\,, \tag{16}$$

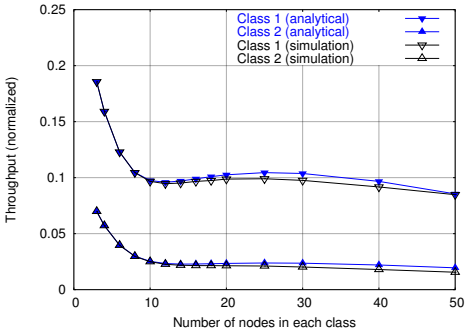
$$\pi_d = A_1 \pi_{d-1} + A_2 \pi_d \tag{17}$$

$$\pi_{S_1} = \pi_{S_1} \frac{1}{CW} + \pi_{C_1} P_{C_2,S_1} + \pi_{C_2} P_{C_2,S_1} + \sum_{k=0}^{d-1} \pi_k P_{k,S_1} + \pi_d P_{d,S_1}\,, \tag{18}$$

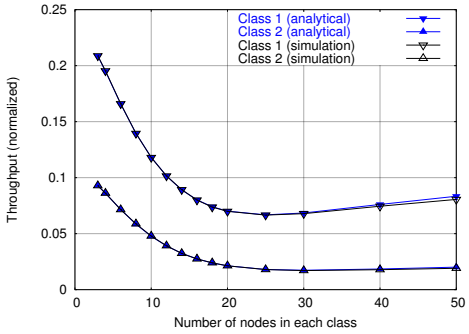
$$\pi_{S_2} = \pi_{S_2} \frac{1}{CW} + \pi_{C_2} P_{C_2,S_2} + \pi_d P_{d,S_2}\,, \tag{19}$$

$$\pi_{C_1} = \sum_{k=0}^{d-1} \pi_k P_{k,C_1} + \pi_{C_1} P_{C_1,C_1}\,, \tag{20}$$

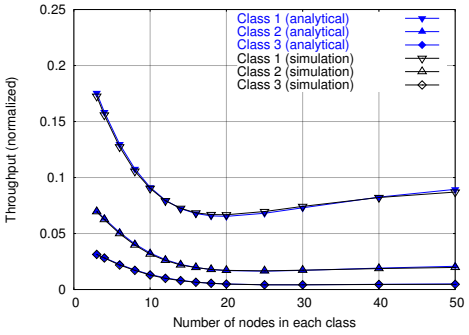
$$\pi_{C_2} = \pi_d P_{d,C_2} + \pi_{C_2} P_{C_2,C_2}\,, \tag{21}$$



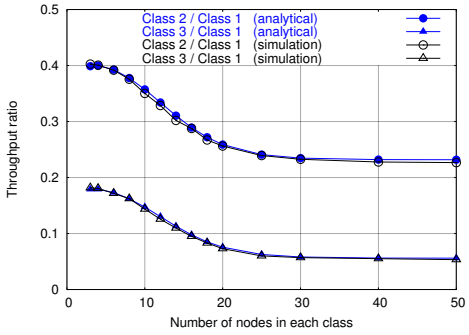
(a) 2 classes: $CW_{1,2}$=[8,8], [16,16]



(b) 2 classes: $CW_{1,2}$=[16,16], [32,32]

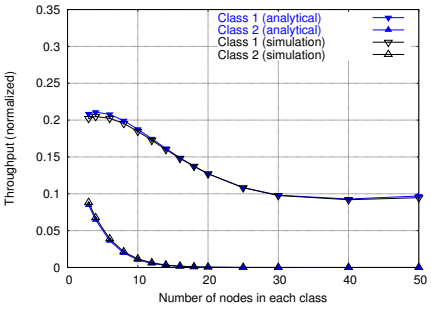


(c-1) 3 classes: $CW_{1,2,3}$=[8,16], [16,32], [32,64]

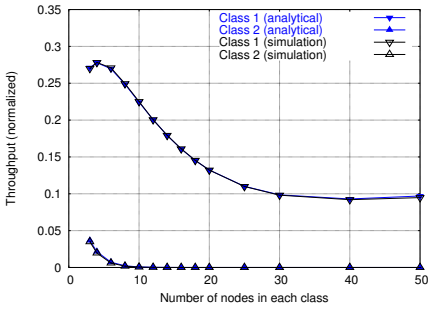


(c-2) Throughput ratios of class 2/1, and 3/1.

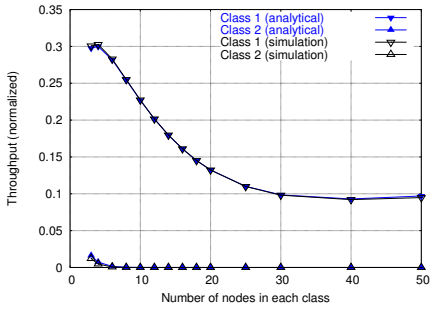
Fig. 5. Effect of CW in EDCA: analytical and simulation results for multiple classes. Each class has a different contention window range but the same AIFS value, $AIFSN = 2$.



(a) 2 classes: $AIFSN_{1,2}$=2, 3



(b) 2 classes: $AIFSN_{1,2}$=2, 5



(c) 2 classes: $AIFSN_{1,2}$=2, 7

Fig. 6. Effect of AIFS in EDCA: analytical and simulation results for multiple classes. Each class has a different AIFS value but the same contention window size, $CW = 16$.

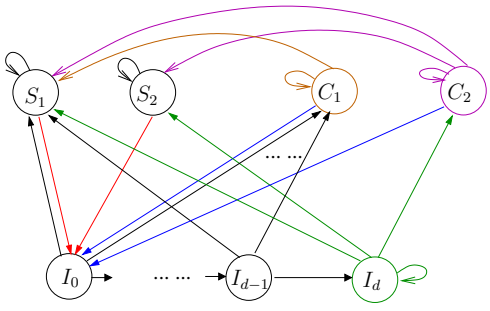


Fig. 7. The discrete Markov chain for the case $AIFS_2 - AIFS_1 = d$ and $CW_1 = CW_2 = CW$.

where $A_1 = (1-\tau)^N$ and $A_2 = A_1^2$. To solve the above series of equations, we have the following observations as the simplified results of Eq. (2)-(10):

$$P_{k,S_1} = N\tau(1-\tau)^{N-1} \tag{22}$$

$$P_{d,S_1} = P_{d,S_2} \tag{23}$$

$$P_{C_2,S_1} = P_{C_2,S_2} \tag{24}$$

Subtracting (19) from (18) results in:

$$\left(1 - \frac{1}{CW}\right)(\pi_{S_1} - \pi_{S_2}) = \pi_{C_1} P_{C_1,S_1} + \sum_{k=0}^{d-1} \pi_k P_{k,S_1} = \pi_{C_1} P_{C_1,S_1} + \frac{N\tau}{1-\tau}\frac{A_1}{1-A_1}(1-{A_1}^d)\pi_0 \tag{25}$$

After a series of simple yet tedious algebra computation, we can finally obtain Δ in the following form:

$$\Delta = \pi_{S_1} - \pi_{S_2} = \frac{1-{A_1}^d}{f_1 + f_2 {A_1}^d}, \tag{26}$$

where $f_1 > 0$ and $f_2 > 0$. They are semi-constants, and can be expressed in terms of N and CW.

$$\frac{\Delta(d+1) - \Delta(d)}{\Delta(d)} = \frac{(1-A_1)(f_1+f_2)}{(f_1 + f_2 {A_1}^{d+1})(1-{A_1}^d)}\ {A_1}^d \tag{27}$$

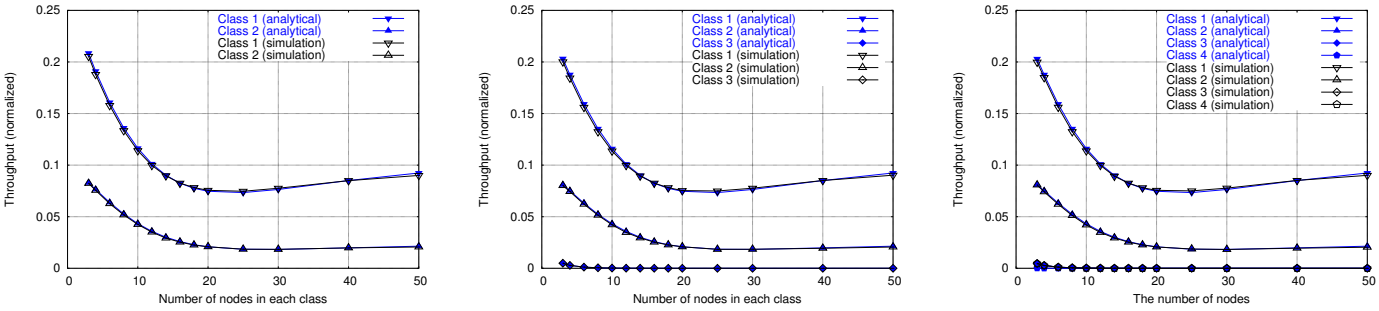
Summarizing these results, the effect of AIFS has the following properties:

Property 1. As $AIFS_2 - AIFS_1 = d$ increases, $\eta_2 - \eta_1 = \Delta$ and consequently the throughput difference increases approximately *exponentially*. The increasing rate depends on $A_1 = (1-\tau)^N$. As the result,

Property 2. As either $N \to \infty$ or $d \to \infty$, $\pi_{S_2} \to 0$.

C. Joint effect of CW and AIFS

Lastly, we study the joint impact of both the contention window and the AIFS value on the performance of service differentiation. We use the default EDCA parameters as specified in Tab. II. Fig. 8 gives both simulation and analytical results for three different combinations of CW and AIFS values. All the observations made in the studies of varying either the contention window size or the AIFS value are observed (although in a mixed manner). In particular, in the presence of stations with the smallest AIFS (AIFSN=2), stations of all the other classes (AC3 and AC4) attain very little (close to zero) throughput.



(a) 2 classes: AC1 and AC2 (b) 3 classes: AC1, AC2 and AC3 (c) 4 classes: AC1-4

Fig. 8. Joint effect of CW and AIFS: analytical and simulation results under different combinations of CW and AIFS values (the default EDCA parameter setting as specified in Tab. II. The configuration of the four access categories, AC1-AC4, are respectively $CW_{1,2,3,4} = [8,16],[16,32],[32,1024],[32,1024]$, $AIFSN_{1,2,3,4} = 2,2,3,7$.

V. Effect of the TXOP Limit

In Section I we have quantitatively explain the different roles that CW and AIFS, and TXOP limit in the EDCA service differentiation mechanism. In general CW and AIFS decide the channel access opportunity, while TXOP decides the channel occupying time after the access is granted. We formally study the effect of the TXOP limit by a theoretical analysis.

Property 1. Approximately, TXOP has the proportional effect on the bandwidth allocation among different classes. In particular, denote the throughput ratio of two classes by r when the two classes have the same TXOP; If we vary TXOP of one class and denote it by $TXOP'$, then the throughput ratio becomes

$$r' \approx r \times \frac{TXOP'}{TXOP}. \tag{28}$$

Proof: Denote the two classes by i and j, respective. From (14) we have

$$r = \frac{\mathbb{P}[\widetilde{s} = S_i]}{\mathbb{P}[\widetilde{s} = S_j]} \tag{29}$$

$$r' = \frac{\eta_i}{\eta_j} = \frac{\overline{m}'}{\overline{m}} \times \frac{\mathbb{P}[\widetilde{s} = S_i]'}{\mathbb{P}[\widetilde{s} = S_j]'} \tag{30}$$

Because the value of TXOP affects the average slot length but does not change the stable probabilities. Therefore, $\mathbb{P}[\widetilde{s} = S_i]' = \mathbb{P}[\widetilde{s} = S_i]$ and $\mathbb{P}[\widetilde{s} = S_j]' = \mathbb{P}[\widetilde{s} = S_j]$. Combined with $\frac{TXOP'}{TXOP} \approx \frac{\overline{m}'}{\overline{m}}$, (28) is trivially true.

Property 2. Consider a network that is operated at the optimal condition specified in Theorem 1 (in Section VI) [1] and denote the total throughput by η_{opt}. Increase the TXOP value of one class will increase the total throughput. That is, denoting the new value of TXOP by TXOP' and the resulted total throughput by η'_{opt},

$$TXOP' > TXOP \Rightarrow \eta'_{opt} > \eta_{opt}. \tag{31}$$

Proof: Given that all nodes use the same TXOP, the total throughput at the optimal condition can be expressed as follows:

$$\eta_{opt} = \frac{\overline{m}\, P_S}{\overline{t}} = \frac{\overline{m}\, P_S}{A_m\, t_s + P_C\, T_C + P_S\, T_S}, \tag{32}$$

where $P_S = \sum_{j=1}^{M} P_{S_j}$, the total successful transmission probability.

Increasing the TXOP value of a class, say, class k, increases the successful transmission time from T_S to T'_S. However, the average collision time, T_C, remains the same. It is because a node will continue to transmit more packets only if it has successfully transmitted the first packet. As the result, the average slot length becomes larger, and so does the payload of class-k's successful transmission. The new throughput becomes

$$\eta'_{opt} = \frac{\overline{m}\, P_S + (\overline{m}' - \overline{m}) P_{S_k}}{A_m\, t_s + P_C\, T_C + P_S\, T_S + P_{S_k}(T'_C - T_C)} = \frac{\overline{m}\, P_S + (\overline{m}' - \overline{m}) P_{S_k}}{\overline{t} + P_{S_k}(T'_C - T_C)} \tag{33}$$

Let $\delta_1 = \overline{m}' - \overline{m}$ and $\delta_2 = T'_C - T_C$. (33) - (32) generates:

$$\eta'_{opt} - \eta_{opt} = \frac{\overline{m}\, P_S P_{S_k} \delta_1}{\overline{t}\,(\overline{t} + P_{S_k}\delta_2)} \times \left(\frac{\overline{t}}{\overline{m}\, P_S} - \frac{\delta_2}{\delta_1} \right) \tag{34}$$

By increasing TXOP, class-k nodes can transmit multiple (say, $Z+1$) data packets in one successful transmission. Then $\delta_1 = Z\overline{m}$ and $\delta_2 = Z(\overline{m} + \epsilon)$, where ϵ = SIFS + (MAC/PHY packet header overhead). $\frac{\delta_2}{\delta_1} = 1 + \frac{\epsilon}{\overline{m}} \approx 1$. On the other hand, $\frac{\overline{t}}{\overline{m}\, P_S}$ is much larger than 1 because the idle and collision states consume non-negligible part of time even in the optimal condition. As the result, $\eta'_{opt} > \eta_{opt}$.

[1] Logically, we shall present Theorem 1 first. However, we feel that it might be more appropriate to put all the discussions on the effect of the three factors in sequence.

VI. A General QoS Provisioning Framework

In this section, we first summarize three important observations obtained from our analytical as well as simulation results presented in Section IV. These results reveal that the default EDCA setting fail to achieve the optimal (in terms of maximal bandwidth utilization) and satisfactory (in terms of meeting user-specified QoS requirement) performance. To address these problems, we first formulate an optimization problem to achieve the goals of maximal bandwidth utilization and proportional bandwidth utilization. And then we propose a general QoS provisioning framework to provide QoS in an optimal way. We also briefly discuss how to incorporate the framework in various network environments.

A. Observations obtained from the analytical and simulation results

From out analytical and simulation results, we have made three insightful observations as follows.

- **Observation 1.** Higher prioritized traffic with smaller AIFS can easily grab most of the bandwidth and starve other traffic. As evidenced by our study presented in Section IV-B, turning AIFS has approximately *exponential* effect on service differentiation. This greatly disfavors the traffic prioritized by larger AIFS and the starvation phenomenon is often not desired.
- **Observation 2.** The bandwidth allocation ratios fail to stay stable and they vary as the network load changes. The results in Section IV-A have shown that tuning CW has the capability of allocation bandwidth among different classes approximately *proportionally*. However, the achieved ratios fail to stay at stable levels and as the number of stations increases they vary significantly. Proportional QoS are usually pre-specified in the form of desired bandwidth allocation ratios. Clearly, the default CW setting in EDCA can not fulfill this QoS performance goal.
- **Observation 3.** The bandwidth is under-utilized. Finally, throughput the results presented in Section IV, we have observed the network throughput (sum of all classes' throughput) generally reduces as the number of stations increases. Similar observations have been made in the case of single class (802.11 DCF), e.g. [2] and [3]. As the number of stations increases, more time is wasted in collisions and the bandwidth is under-utilized. Previous work on 802.11 DCF has shown the existence of an optimal point that can achieve maximal bandwidth utilization. We will show the existence of such an optimal point and identified the optimality condition in the multiple case later in the next section.

B. Achieving maximal bandwidth utilization and proportional bandwidth allocation

As mentioned above, schemes that assign small AIFS values to high-priority access categories risk the possibility that stations of low-priority access categories (AC3–AC4) will starve. As such, we propose to consider only the dimension of varying contention window sizes in provisioning deterministic QoS to best-effort traffic.

For most network applications, the throughput attained by stations of different classes is perhaps the major measure of quality of service. As the *available* bandwidth in a wireless environment is variable and changes as the number of stations increases, instead of providing QoS in the form of absolute bandwidth, we aim to provide deterministic *proportional* QoS among best-effort traffic. We define the ratio, r_j, $j = 2, \cdots, M$, of per-station throughput attained by a station of class j to that attained by a station of class 1, i.e., $\frac{\eta_j}{N_j} = r_j \frac{\eta_1}{N_1}$, $j = 2, \cdots, M$.

As indicated in Section IV, given the parameter settings currently recommended in [13], [17], the throughput ratio r_j dynamically changes as the number of stations varies (Fig. 5(c-1)) and cannot be fixed at a stable level. Also, it is not clear whether or not the current parameter setting renders the maximal system throughput, although it has been proved in [3] that the current IEEE 802.11 parameter setting cannot achieve the maximal system throughput. In what follows, we study, by leveraging the analytical model derived in Section III, how the contention window sizes can be optimally set to provide deterministic proportional QoS and to maximize the system throughput.

Formulation of the throughput maximization and deterministic proportional QoS provisioning problem: We consider a single-cell network with M classes. Stations of all the classes are configured with the same AIFS value, but are assigned different contention window sizes CW_j, $1 \leq j \leq M$. Then the problem of combined throughput maximization and deterministic proportional QoS provisioning can be formally stated as

Problem 1: Given the throughput ratio r_j, $j = 2, \cdots, M$, determine the optimal contention window sizes CW_j, $j = 1, \cdots, M$ such that

$$\text{Maximize } \eta = \sum_{j=1}^{M} \eta_j \tag{35}$$

$$\text{s.t. } \frac{\eta_j}{N_j} = r_j \frac{\eta_1}{N_1}, \text{ for } j = 2, \cdots, M \tag{36}$$

Proposed solution: In the model given in Section III, the system throughput is derived as a function of the number of class-i stations (N_i), the contention window size (CW_i), the AIFS values (a_i). In principle, the optimization problem can be solved numerically. However, a simple, closed-form solution would be desirable so that stations can dynamically track the parameters in the solution, and on-line calculate the optimal solution.

Before delving into the derivation, we make the following observation. Figure 9 depicts the system throughput as a function of the contention window size (CW) in the case of one traffic class ($N = 5, 10, 20$ and 50). As shown in Fig. 9 (and as mentioned earlier), the analytical results derived under the proposed model agree very well with the simulation results, but those derived under the p-persistent model (which assume that all the stations independently access to *any* slot with a fixed probability) fails to do so. Nevertheless, both models give approximately the same optimal value of CW at which the system throughput is maximized. This is not a coincidence, because at the operational point where the maximal throughput is achieved (e.g., $CW = 20$ when $N = 5$), the channel is not overly congested and the peculiar effect of the access pattern to the *post-busy* slot has not yet became significant. Similar trends have also been observed in the case of multiple traffic classes. This observation suggests that as far as derivation of the optimal congestion window size is concerned, one can leverage the p-persistent model, subject to the proportional constraint Eq. (36).

In the p-persistent model, stations of class j transmit in *a* slot independently and uniformly with probability τ_j. Given the contention window size CW_j, τ_j can be calculated as $\tau_j = \frac{2}{CW_j+1}$. Then the stationary probabilities of channel states, i.e., the *idle* state, the *successful class-j transmission* state, and the *collision* state, can be readily derived as

$$P_I = \prod_{j=1} (1-\tau_j)^{N_j} = A_M, \tag{37}$$

$$P_{S_j} = N_j \frac{\tau_j}{1-\tau_j} A_M, \tag{38}$$

$$P_C = 1 - P_I - \sum_{j=1}^{M} P_{S_j}. \tag{39}$$

For ease of exposition, we assume that the size of all data frames is a constant, and thus the duration of a successful transmission and a collision period are the same (i.e., $T_D = T_C$). (The assumption can be relaxed with modest modification.) Plugging the above stationary probabilities into Eqs. (13) and (14), we have

$$\eta_j = \frac{\overline{m}}{T_D} \frac{N_j \frac{\tau_j}{1-\tau_j} A_M}{1 - A_M \left(1 - \frac{t_s}{T_D}\right)}. \tag{40}$$

Now we are in a position to derive the optimal value of CW_j, $1 \le j \le M$.

Theorem 1: **(Optimality Condition)** Given the expression for system throughput, Eq. (40), the optimal solution to Problem 1 (defined in Eqs. (35) and (36)) is: for $j = 1, \cdots M$,

$$CW_j^* = \frac{\sqrt{2\beta T_D'}}{r_j} + 1, \tag{41}$$

where $\beta = \left(\sum_{j=1}^{M} N_j\, r_j\right)^2 - \sum_{j=1}^{M} N_j\, r_j^2$ and $T_D' = T_D/t_s$, that is, the duration of a successful transmission in the unit of slots. $r_1 \equiv 1$.

Proof: Refer to Appendix IX.

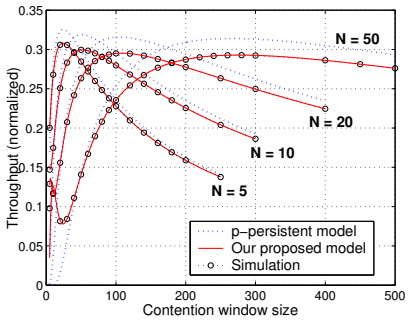
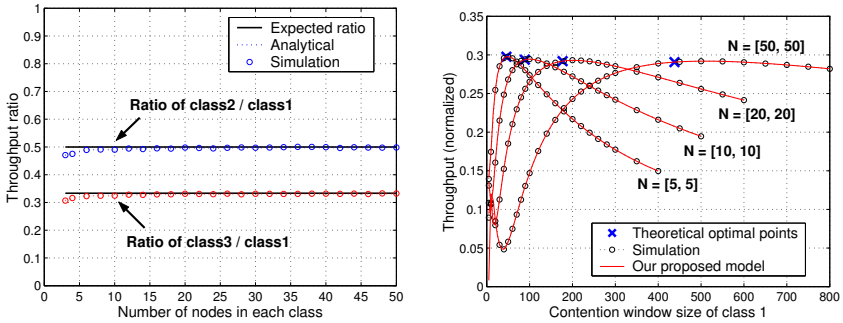


Fig. 9. The relationship between the saturation throughput and the contention window size. There is only one class in the system. N is the number of nodes.



(a) (b)

Fig. 10. The throughput ratio among different traffic classes ((a)) and the system throughput ((b)) in the case that there are 3 traffic classes and the proposed solution (given in Eq. (41)) is used to calculate the optimal contention window size. The QoS specified is $[r_2, r_3] = [0.5, 0.33]$.

Discussion: Theorem 1 gives the optimality condition satisfying which the network can achieve our goals. However, how to realize it largely depends on the specific network configuration and the means may vary in a large spectrum. This is out of the scope of this report but we have conducted some study in this avenue, for example, the one reported in [7].

Throughout the discussion, we have assumed that the basic access method without the RTS-CTS floor acquisition mechanism is used, and that all stations operate at a common data rate. However, the results can be readily extended to accommodate more general scenarios in which such assumptions are relaxed.

C. A general QoS provisioning framework

To achieve our goals of maximal bandwidth utilization and user-specified QoS performance, we devise a general QoS framework that leverage the EDCA service differentiation mechanism judiciously to provide QoS in an optimal way.

- Suggested by Observation 1, the means of using small AIFS to achieve fast channel access has to be carefully used to avoid burst contention. Traffic with small AIFS has to be carefully controlled in their intensity. Take periodic real-time traffic as example, we suggest that they only use small AIFS to obtain admission and to make reservation; following that, they access the channel using reservation-based access. The design of the reservation schedule shall be adapted for specific types of networks.
- Also suggested by Observation 1, normal traffic uses contention-based access and are configured with the same AIFS. Here normal traffic mainly refers to best-effort traffic for data transfer. QoS for flows of this type of traffic is often specified in the form proportional bandwidth allocation. The means of assigning different CW values/ranges is appropriate and thus chosen for this purpose.
- Indicated by Observation 2 and 3, the network shall be operated at the *optimal condition* derived in Section VI-B to achieve the maximal bandwidth utilization and deterministic proportional bandwidth allocation.

Remark: The proposed QoS framework is a set of theoretically grounded guidelines, rather than a detailed algorithm. As we pointed out at the beginning of the section, the realization means can vary from one network to another. In our work reported in [7], we show that the proposed QoS provisioning framework can be readily incorporated in the UWB WPANs and greatly enhance its network performance. Applying the framework in different environments, such as WLANs with AP support and wireless mesh networks, are of our future research interest.

VII. Conclusion

In this report, we have conducted a rigorous, comprehensive, and theoretical analysis of IEEE 802.11e EDCA, and have shown that with the currently recommended parameter setting, EDCA cannot provide satisfactory QoS. In

particular, both our analytical and simulation results have shown that 1) stations of a high-priority class (i.e., with a small AIFS value) will dominate the channel access, depriving stations of the other classes the chance to access the channel. And 2) without responding to the system dynamics (e.g., taking into account of the number of active class-i stations), EDCA cannot allocate bandwidth in a deterministic proportional manner and the system bandwidth is under-utilized.

After identifying the deficiency of EDCA in service differentiation, we propose a general QoS framework to provide QoS in an optimal way. In this framework, 1) real-time traffic use a *contention-based* reservation access method; 2) best-effort traffic is provided with deterministic proportional QoS; and moreover, 3) the bandwidth utilization is maximized. How to incorporate this framework in various network environments, is our next research avenue.

REFERENCES

[1] MultiBand OFDM Alliance SIG, MultiBand OFDM physical layer proposal for IEEE 802.15 task group 3a, Sept. 2004.

[2] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE JSAC*, 18(3), Mar. 2000.

[3] F. Cali, M. Conti, and E. Gregori. Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit. *IEEE/ACM Trans. on Networking*, 8(6), Dec. 2000.

[4] S. Choi, J. del Prado, S. Shankar, and S. Mangold. IEEE 802.11e contention-based channel access (EDCF) performance evaluation. In *Proc. of IEEE ICC*, May 2003.

[5] Y. Ge. QoS provisioning for IEEE 802.11 MAC protocols. Ph.D Thesis, University of Ohio State, 2004.

[6] C. Hu, H. Kim, and J. C. Hou. An analysis of the binary exponential backoff algorithm in distributed MAC protocols. In *Tech. Rep. No. UIUCDCS-R-2005-2599. http://lion.cs.uiuc.edu/∼chunyuhu*, July 2005.

[7] C. Hu, H. Kim, J. C. Hou, D. Chi, and S. S. Nandagopalan. Provisioning quality controlled medium access in ultrawideband WPANs. In *Proc. of IEEE INFOCOM*, April 2006.

[8] J. Hui and M. Devetsikiotis. Performance analysis of IEEE 802.11e EDCA by a unified model. In *Proc. of GLOBECOM*, Nov. 2004.

[9] B. Li and R. Battiti. Performance analysis of an enhanced IEEE 802.11 distributed coordination function supporting service differentiation. In *Quality for All, QofIS*, 2003.

[10] A. Lindgren, A. Almquist, and O. Schelen. Evaluation of quality of Service schemes for IEEE 802.11 wireless LANs. In *Proc. of IEEE LCN*, Nov. 2001.

[11] A. Lindgren, A. Almquist, and O. Schelen. Quality of service schemes for IEEE 802.11 wireless lans - an evaluation. *In Special Issue of the Journal on Special Topics in MONET on Performance Evaluation of QoS Architectures in Mobile Networks*, 8(3), June 2003.

[12] S. Mangold, S. Choi, P. May, and G. Hiertz. IEEE 802.11e - fair resource sharing between overlapping basic service sets. In *Proc. of IEEE PIMRC*, Sept. 2002.

[13] MBOA. Distributed medium access control (MAC) for wireless networks. Draft specification 0.93, Feb. 2005.

[14] D. Pong and T. Moors. Call admission control for IEEE 802.11 contention access mechanism. In *Proc. of IEEE GLOBECOM*, Dec. 2003.

[15] V. Ramaiyan and A. Kumar. Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness, multistability and throughput differentiation. In *Proc. of ACM SIGMETRICS*, June 2005.

[16] J. W. Robinson and T. S. Randhawa. Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function. *IEEE JSAC*, 22(5), 2004.

[17] J. D. Sai Shankar N, V. Gaddam and K. Challapali. The new MBOA MAC specification: A distributed MAC protocol and OFDM based PHY for next generation WPANs.

[18] S. Shankar, J. del Prado Pavòn, V. G, and K. Challapali. Performance evaluation of the multiband OFDM alliance (MBOA) specification: A distributed MAC protocol and OFDM PHY layer for next generation ultra wide band (UWB) WPANs. In *submission*, 2005.

[19] Y. Xiao. An analysis for differentiated service in IEEE 802.11 and IEEE 802.11e wireless LANs. In *Proc. of IEEE ICDCS*, Mar. 2004.

[20] J. Zhao, Z. Guo, Q. Zhang, and W. Zhu. Performance study of MAC for service differentiation in IEEE 802.11. In *Proc. of GLOBECOM*, 2002.

VIII. LEMMA 1 AND LEMMA 2

Lemma 1: Given the Markov chain described in Section III (Fig. 4), the transition probability from state C_{a_j}, $j = m, \cdots, M$ to state I_{a_1} can be approximately expressed as

$$P[C_{a_j} \rightarrow I_{a_1}] = \frac{1}{P[I_k \rightarrow C_{a_j}]} \left\{ B_m - A_j \left[1 + \sum_{k=1}^{m} \frac{N_k \tau_k}{1-\tau_k} \left(1 - \frac{1}{CW_k} \right) \; + \sum_{k=m+1}^{j} \frac{N_k \tau_k}{1-\tau_k} \right] \right\}. \tag{42}$$

Proof: $P[C_{a_j} \rightarrow I_{a_1}]$ is the probability that the *post-busy* slot after a collision slot C_{a_j} is idle. For notational convenience, tag the collision slot and its *post-busy* slot by $slot_1$ and $slot_2$, respectively. Let E_j denote the event that $(n_1, n_2, \cdots, n_j)$ stations transmit in $slot_1$, where n_k ($n_k = 0, 1, \cdots, N_k$) is the number of stations in class k. We have

$$\begin{aligned} P[C_{a_j} \rightarrow I_{a_1}] &= P\left[slot_2 \text{ is idle, given } slot_1 \text{ is a collision of } C_{a_j} \right] \\ &= \underbrace{\sum_{n_1} \sum_{n_2} \cdots \sum_{n_j}}_{\sum_{k=1}^{j} n_k \geq 2} P\left[slot_2 \text{ is idle } | E_j \right] \; P\left[E_j | \; slot_1 \text{ is } C_{a_j} \right]. \end{aligned} \tag{43}$$

Given the event E_j, $slot_2$ is idle if and only if none of the stations of the first m classes transmit in $slot_2$. Therefore,

$$P\left[slot_2 \text{ is idle } | E_j \right] = \prod_{k=1}^{m} \left(1 - \frac{1}{CW_k} \right)^{n_k}. \tag{44}$$

To compute precisely the probability $P\left(E_j | slot_1 \text{ is } C_{a_j} \right)$ it is necessary to enumerate the channel states and expand the channel state space $\mathbb{S}$ into $(n_1, n_2, \cdots, n_M)$. Hence, we leverage the fact that the channel state immediately following a busy period is mostly likely to be the idle state, because only nodes that are involved in the busy period will contend in the *post-busy* slot ($slot_2$). Based on this argument (which was collaborated by the simulation results), we have

$$P\left[E_j | \; slot_1 \text{ is } C_{a_j} \right] = \frac{1}{P[I_k \rightarrow C_{a_j}]} \prod_{k=1}^{j} \binom{N_k}{n_k} {\tau_k}^{n_k} (1-\tau_k)^{N_k - n_k}. \tag{45}$$

Plugging Eqs. (44) and (45) into Eq. (43), and after performing some algebraic operations, (42) follows. ∎

Lemma 2: Given the Markov chain described in Section III (Fig. 4), the transition probability from state C_{a_j}, $j = m, \cdots, M$ to state S_u ($u \leq m$) can be approximately expressed as

$$P\left[C_{a_j} \rightarrow S_u \right] = \frac{1}{P[I_k \rightarrow C_{a_j}]} \; N_u \; \frac{\tau_u}{CW_u} \left(\frac{B_m}{1 - \frac{\tau_u}{CW_u}} - \frac{A_j}{\tau_u} \right). \tag{46}$$

Proof: Following the notation defined in the proof of Lemma 1, $P[C_{a_j} \rightarrow S_u]$ is the probability $slot_2$ is a successful transmission of class u, given that $slot_1$ is a collision C_{a_j}; or simply $P[slot_2 \text{ is } S_u | slot_1 \text{ is } C_{a_j}]$. By conditioning on the event E_j, we have

$$P\left[C_{a_j} \rightarrow S_u \right] = \underbrace{\sum_{n_1} \sum_{n_2} \cdots \sum_{n_j}}_{\sum_{k=1}^{j} n_k \geq 2} P\left[slot_2 \text{ is } S_u | E_j \right] \; P\left[E_j | \; slot_2 \text{ is } C_{a_j} \right]. \tag{47}$$

A successful transmission of class u follows a collision of C_{a_j} if and only if only one station of class u chooses to transmit in $slot_2$ with probability $\frac{1}{CW_u}$. That is,

$$\begin{aligned} & P\left[slot_2 \text{ is } S_u | E_j \right] \\ &= n_u \frac{1}{CW_u} \left(1 - \frac{1}{CW_u} \right)^{n_k - 1} \prod_{k=1, k \neq u}^{m} \left(1 - \frac{1}{CW_k} \right)^{n_k} \\ &= \frac{n_u}{CW_u - 1} \prod_{k=1}^{m} \left(1 - \frac{1}{CW_k} \right)^{n_u}. \end{aligned} \tag{48}$$

The probability $P\left[E_j | slot_1 \text{ is } C_{a_j} \right]$ is the same as Eq. (45). Plugging Eqs. (48) and (45) into Eq. (47), we can derive $P\left[C_{a_j} \rightarrow S_u \right]$ as given in Eq. (46). ∎

IX. Proof of Theorem 1

Theorem 1: Given the expression for system throughput, Eq. (40), the optimal solution to Problem 1 (defined in Eqs. (35) and (36)) is: for $j = 1, \cdots M$,

$$CW_j^* = \frac{\sqrt{2\beta T_D'}}{r_j} + 1, \tag{49}$$

where $\beta = \left(\sum_{j=1}^{M} N_j\, r_j\right)^2 - \sum_{j=1}^{M} N_j\, r_j^2$ and $T_D' = T_D/t_s$, that is, the duration of a successful transmission in the unit of slots. $r_1 \equiv 1$.

Proof: The throughput by class j (Eq. (40)) is

$$\eta_j = \frac{\overline{m}}{T_D}\, \frac{N_j \frac{\tau_j}{1-\tau_j} A_M}{1 - A_M \left(1 - \frac{t_s}{T_D}\right)},$$

where $\overline{m}$, t_s and T_D are constant variables, representing the average data frame payload, the length of an idle slot and the duration of a successful transmission, respectively. $A_M = \prod_{h=1}^{M}(1-\tau_h)^{N_h}$ as defined in Eq. (2).

Let $x \triangleq \frac{\tau_1}{1-\tau_1}$, $c_1 \triangleq \frac{\overline{m}}{T_D}$, and $c_2 = 1 - \frac{t_s}{T_D}$ and $A \triangleq A_M$. Then we can further simplify the above equation as

$$\eta_j = c_1 \frac{N_j \frac{\tau_j}{1-\tau_j} A}{1 - c_2 A}. \tag{50}$$

To fulfill the proportional bandwidth allocation requirement Eq. (36), we have

$$\begin{aligned} r_j = \frac{\eta_j}{N_j} / \frac{\eta_1}{N_1} &= \frac{\tau_j}{1-\tau_j}\, \frac{1}{x} \\ \Rightarrow \tau_j &= \frac{r_j\, x}{1 + r_j\, x}. \end{aligned} \tag{51}$$

The total system throughput is the summation of the throughput achieved by each class, i.e.,

$$\begin{aligned} \eta = \sum_{j=1}^{M} \eta_j &= c_1 \sum_{j=1}^{M} N_j\, r_j \frac{xA}{1 - c_2 A} \\ &= \left(c_1 \sum_{j=1}^{M} N_j\, r_j\right) \frac{x}{\frac{1}{A} - c_2}. \end{aligned} \tag{52}$$

By Eq. (51), we have $1 - \tau_j = 1 - \frac{r_j\, x}{1 + r_j\, x} = \frac{1}{1 + r_j\, x}$, and

$$A = \prod_{j=1}^{M} (1-\tau_j)^{N_j} = \frac{1}{\prod_{j=1}^{M} (1 + r_j\, x)^{N_j}}. \tag{53}$$

Using the Taylor series to approximate A, we have

$$\begin{aligned} (A^{-1})(x) &= 1 + \theta\, x + \frac{\beta}{2!}\, x^2 + o(x^2) \\ &\approx 1 + \theta\, x + \frac{1}{2} \beta\, x^2, \end{aligned} \tag{54}$$

where

$$\theta = (A^{-1})'(0) = \sum_{j=1}^{M} N_j\, r_j, \tag{55}$$

$$\beta = (A^{-1})''(0) = \left(\sum_{j=1}^{M} N_j\, r_j\right)^2 - \sum_{j=1}^{M} N_j\, r_j^2. \tag{56}$$

The total system throughput can then be expressed in terms of x:

$$\eta(x) = c_1\,\theta \frac{x}{1+\theta x+\frac{1}{2}\beta x^2 - c_2} = \frac{c_1\,\theta}{\frac{1-c_2}{x}+\theta+\frac{1}{2}\beta x}. \tag{57}$$

It is easy to see $\eta(x)$ is maximized at $x^* = \sqrt{\frac{1-c_2}{\beta/2}} = \sqrt{\frac{2}{\beta T_D'}}$, where $T_D' \triangleq \frac{T_D}{t_s}$.

From the two relations, $\tau_1 = \frac{2}{CW_1+1}$ and $x = \frac{\tau_1}{1-\tau_1}$, it is easy to obtain $CW_1 = \frac{2}{x}+1$. Therefore, the system throughput $\eta(x)$ is maximized at

$$CW_1^* = \sqrt{2\beta T_D'} + 1, \tag{58}$$

and the proportional bandwidth allocation is achieved at

$$CW_j^* = \frac{\sqrt{2\beta T_D'}}{r_j} + 1, \tag{59}$$

for $j = 2, \cdots, M$. ■