

Performance analysis of differentiated QoS in IEEE 802.11e WLANs[¶]

Kenan Xu^{1,‡}, Quanhong Wang^{1,§} and Hossam Hassanein^{2,*†}

¹ *Department of Electrical, Computer Engineering, Queen's University, Kingston, Ont., Canada K7L 3N6*

² *School of Computing, Queen's University, Kingston, Ont., Canada K7L 3N6*

SUMMARY

This paper proposes a multi-dimensional Markov model to analyse the performance of the IEEE 802.11e EDCF MAC protocol. Based on this model, we present extensive performance evaluation in terms of throughput, throughput ratios, and access delay of flows of distinct priorities under RTS/CTS mode. We also provide quantitative analysis of the impact of prioritized parameters, i.e. Arbitration InterFrame Space (AIFS), Contention Window (CW) on Quality of Service (QoS) differentiation. The accuracy of the proposed model is verified by means of comparing the numerical results obtained from both analytical model and simulations. Our research can be used as a guideline for the prediction of how flows belonging to a certain Traffic Category (TC) perform with their TC-specific parameters, as well as designing EDCF-based WLANs and tuning the parameters to achieve the desirable differentiated QoS objectives. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Quality of service (QoS); service differentiation; MAC; IEEE802.11e; EDCF; WLANs; Markov chain

1. INTRODUCTION

The Enhanced Distributed Coordination Function (EDCF) is a Quality of Service (QoS)-enabled multiple access scheme defined by IEEE802.11E standard draft [1]. It is a fully distributed, CSMA/CA-based MAC algorithm, but provides access to the Wireless Media (WM) with up to eight priorities, also known as Traffic Categories (TCs). The differentiated access is achieved by having each QoS station (QSTA) operate maximum eight output queues, which are called Virtual Stations (VSs); each VS independently contends for channel access.

*Correspondence to: Hossam Hassanein, School of Computing, Queen's University, Kingston, Ont., Canada K7L 3N6.

† E-mail: hossam@cs.queensu.ca

‡ E-mail: xu@cs.queensu.ca

§ E-mail: qwang@ee.queensu.ca

¶ An abbreviated version of this paper appeared in IEEE Globecom 2003.

Contract/grant sponsor: Communications and Information Technology Ontario (CITO)

Contract/grant sponsor: Natural Science and Engineering Research Council of Canada (NSERC)

Received 7 April 2004

Revised 30 October 2004

Accepted 20 November 2004

To access WM, VSs in QSTAs execute 'listen-before-talk' scheme. Before transmission, a VS will keep sensing the channel until it is detected idle for a period of time, denoted by arbitration interframe space (AIFS). To reduce the probability of collision, the STA will defer its transmission for a random backoff time, which is represented by an integer random backoff counter, and chosen between $[1, CW + 1]$, where CW denotes contention window. The initial CW is set to be CW_{\min} , the lower bound of CW . In case of a collision, CW will be enlarged as $CW_{\text{new}} \geq ((CW_{\text{old}} + 1) * PF) - 1$, where PF denotes persistence factor, which is a value between $[15, 27]$. CW keeps growing with consecutive collisions, until it reaches CW_{\max} , it will remain at this value until it is reset to CW_{\min} upon a successful transmission. The VS is not allowed to transmit until the backoff counter reaches zero. The channel is monitored continuously, the backoff counter will decrement by 1 for each idle slot time; if the channel is sensed busy, countdown of backoff counter will be suspended immediately until the channel is sensed idle for another AIFS. If two VSs finish the countdown at the same time, a virtual collision would occur inside the node. In such case, lower priority packet should yield this TXOP to high priority packet. As a consequence, the lower priority VS will proceed by updating its CW as a collision has happened, and will wait for the next TXOP. All these parameters, including AIFS, CW_{\min} , CW_{\max} , PF , are TC-specific, so that each VS will experience differentiated delivery QoS (e.g. bandwidth and delay).

Moreover, to reduce the duration of a collision when long messages are transmitted, EDCF employs Request-To-Send(RTS)/Clear-To-Send(CTS) based four way hand-shaking algorithm. Under RTS/CTS mode, before transmitting a packet, VS reserves the channel by exchanging RTS/CTS messages, and upon reception of a successfully transmitted packet, the destination VS will send an ACK. Both CTS and ACK are transmitted immediately after the channel idle for a period of time called short interframe space (SIFS), which is shorter than a AIFS. Therefore, no other VSs could have the opportunity to send any other packets during the reserved period. Figure 1 shows an example of EDCF operation with two VSs belonging to two different TCs.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work on distributed QoS-enabled MAC protocol for WLANs, which is followed by our research motivation. In Section 3, we define our multi-dimensional Markov chain model and derive the

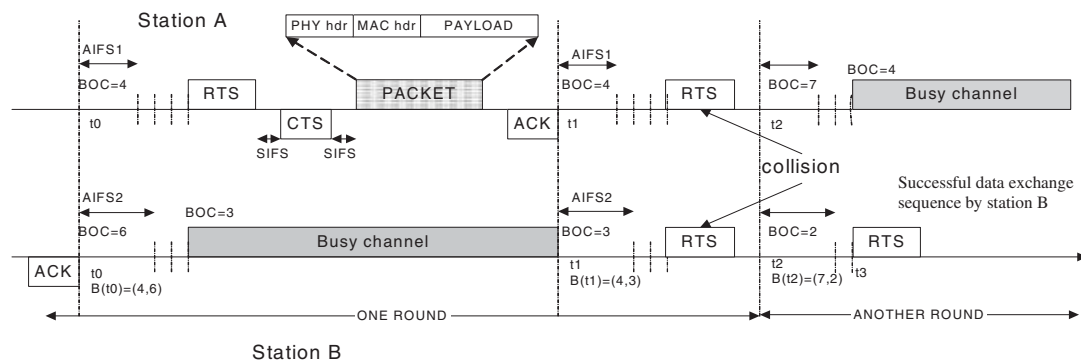


Figure 1. Example of the operation of the EDCF function.

formulation for throughput and delay. In Section 4, we validate the results of our model by comparing with simulation results. In Section 5, we carry out the performance evaluation, with insights of how much the parameters of EDCF impact differentiation QoS to different TCs. Finally, in Section 6, we conclude the paper.

2. RELATED WORK AND MOTIVATION

Besides EDCF specified by task group E, there are other efforts to enhance the CSMA/CA access scheme to provide QoS. Research in References [2–15] has revealed that, based on CSMA/CA access scheme, packet-level QoS can be achieved by differentiating the operation parameters of STAs, enhancing the CW algorithm and separating traffic flows. MACAW in Reference [2] was one of an early MAC protocol intended to provide QoS, particularly per flow fairness. Among other mechanisms, multiplicative increase/linear decrease (MILD) CW algorithms, ‘backoff copy’, and per-flow CW paradigm were proposed to provide better fair access to packet streams across multiple cells.

In Reference [3], efficient channel utilization and weighted fairness were achieved by dynamically selecting the proper CW to reflect the relative weights among data traffic flows and the number of the stations contending for the wireless medium. Therefore, the window size has to be calculated based on the network traffic condition, which is monitored and tracked down along with the operation. The work in Reference [4] took the similar approach to achieving per-flow fair medium access.

In References [5, 6], balanced media access methods (BMAM) was proposed to provide per-flow fair transmission. BMAM introduced the concept of persistence factor, say p , as the leverage of fairness, which is similar to p -persistence CSMA. In BMAM, stations send packets with probability p , after the backoff period, or back off again with probability $1 - p$ using the same CW value, where p can be calculated dynamically using either contention- or time-based media access method, in order to achieve fairness objective.

In Reference [7], a general analytical framework was introduced to model system wide per-flow fairness via the specification of per-flow utility functions. Then the fairness model could be translated into a corresponding contention resolution algorithm. Using this translation, the backoff algorithm for achieving proportional fairness was derived.

The work in References [8, 9] presented a set of cooperative mechanisms that compose a fully distributed wireless differentiated services network. Among others, a distributed, differentiated services-capable MAC was described. The protocol is a modification of DCF in IEEE802.11 and data transmission follows the same listen-before-talk scheme. However, best effort and premium traffic streams use the different CW bounds so that premium traffic tends to wait shorter than the best effort traffic before initiating data exchange sequence. The work in Reference [10] takes the similar approach to enabling QoS in MAC layer.

In Reference [11], three DCF factors, namely, CW size, interframe spaces and packet lengths were studied as the leverage of service differentiation. The authors simulated and compared the delay and bandwidth of prioritized flows when using one of three differentiation mechanisms. Firstly, each priority level has a different backoff increment function. Secondly, each priority level is assigned a different DIFS length. And lastly, each priority level has a different maximum frame length, beyond which the frame has to be segmented and transmitted multiple times.

Similarly, Reference [12] proposed and evaluated different CW algorithms for the prioritized flows.

The work in References [13,14] attempted to achieve weighted fairness of flows with distributed control, which is facilitated by transmission status information exchange. The scheme introduced in Reference [13] tries to mimic the centralized Self-Clocked Fair Queuing (SCFQ) algorithm using the piggybacked timer information and selecting backoff intervals proportional to normalized packet sizes by flow weights, i.e. packet size/weight. Paper [14] chose a CW size proportional to normalized rate of each flow.

In terms of EDCF performance on QoS support, we are on the very early stage of research. A few research efforts have been reported in the literature [16–22]. The work in Reference [16] provided a brief illustration of differentiated QoS effect of EDCF function with simulation results. Reference [17] presented a simulation study of IEEE802.11e in the more realistic scenarios. Reference [18] evaluated and compared four QoS support WLAN MAC schemes, including EDCF, using the simulations. References [19,20] presented an adaptive service differentiation scheme called adaptive EDCF. This scheme is derived from EDCF, by incorporating a dynamic CW algorithm, which took into account both application requirements and network conditions. The performances of EDCF and AEDCF algorithms were simulated and compared with prioritized traffic using simulation. Reference [21] proposed to enhance EDCF with a dynamic traffic class management protocols and provided a sliding QoS differentiation mechanism among traffic classes to cope with the instantaneous channel fluctuations. However, little mathematical analysis on EDCF has been reported, partially because it is a very challenging task. Reference [22] derived an analytical model for the throughput performance of an 802.11e Wireless LAN under the EDCF, and proposed an admission control and parameter configuration algorithm that provides the committed throughput guarantees and accepts as many requests as possible. However, the model failed to consider the mutual impact of traffic flows of different classes.

EDCF operation is a complex and high-dynamic process; the performance is affected by a number of factors. Intuitively, the smaller AIFS is, the shorter the waiting period encountered before a VS accessing to the medium. Thus the VS obtains the higher priority when contending for the WM. And this is the same as what the smaller CW does. However, AIFSs differentiation results in more complicated effect in term of the degree of differentiation, which not only depends on the length of AIFSs, but also relies on the combination of AIFS and CW. In this paper, we consider the combined impact of AIFSs and CW related parameters and provide quantitative analysis of QoS supported by EDCF in terms of Saturation Throughput (ST)^{||} and access delay by means of a multi-dimensional Markov model. We derive formulas to compute the systems ST, ST ratios and access delay of different flows as function of AIFS and CW parameters. We validate our theoretical analysis by comparing with simulation results. These lead to an accurate evaluation of differentiated QoS provided by IEEE802.11e EDCF as well as the analysis of the impacts of the prioritized parameters on QoS differentiation. Based on these results, we obtain guidelines for the prediction of how the prioritized VSs perform with specific parameters, as well as how to tune the parameters to achieve the desired QoS objectives.

^{||}Saturation throughput is defined as the limit reached by the system throughput as the offered load increases till the maximum load that the system can carry in stable conditions. This concept is also interpreted in Reference [14].

3. ANALYTICAL MODEL

We focus on the differentiated QoS provided by the IEEE802.11e EDCF under overload conditions. Indeed, the major contributions of this paper are the analytical derivation of system ST, throughput and access delay of individual flows, and the accurate calculation of throughput ratios among flows. The analytical results describe the pattern in which prioritized flows share. More importantly, the calculation of throughput ratios leads to quantitative understanding of to what degree High-Priority (HP) flows can have an advantage over Low-Priority (LP) ones when contending for the WM. Therefore, we gain insight on how well 802.11e EDCF can actually support differentiated QoS using VSs with different AIFS lengths and CW related parameters.

The rest of this section consists of three parts. First, we present and discuss the necessary assumptions used in our analysis. Then, we present an m -dimensional Markov chain model, and define the state space as well as the transition matrix. Last, according to the theorem of transient states and time to absorption [23, 24], we derive formulation of throughput, throughput ratios among different flows, and the access delay as functions of AIFS lengths and the CW size.

3.1. Model assumptions

A wireless LAN executing EDCF function is a complex system. The performance is influenced by many factors [25]. It is very difficult to take every factor into account at the same time. We make a few necessary assumptions in order to simplify the modelling of system without losing the major operational characteristics of the system.

First, in this paper, we study the EDCF performance in single-hop wireless LANs. We assume the network is fully connected, i.e. every node in the system can hear all other nodes directly. In this case, hidden and exposed terminals would not arise.

Second, we assume ideal channel condition. We ignore the packet loss due to interference of all kinds or capture effects. Thus only packet loss due to packet collision is considered. Packet collision happens when more than one node transmits at the same time.

Third, we analyse EDCF performance when the system operates under saturation conditions, i.e. each VS always has a packet available for transmission; in other words, the transmission queues of the station are always non-empty. While this is not always the case in practice, we should note that the 'ST' is a fundamental performance metric defined as the limit reached by the system throughput as the offered load increases, and represents the maximum load that the system can carry in stable conditions. Similar to other random access schemes, CSMA/CA access methods exhibit a non-linearity in term of throughput change when the traffic load increase to a certain degree. Reference [26] has illustrated and discussed this characteristic of DCF. As the offered load increases, the throughput grows up to a maximum value, referred to as maximum throughput. However, further increase of the offered load leads to a decrease in the system throughput. This results in the practical impossibility to operate DCF function at its maximum throughput for a 'long' period of time. ST indicates the throughput lower bound when the network is running under high load. As well, the access delay derived in saturation condition gives the upper bound for average packet service time. The access delay refers to the time interval between a packet becoming Head of Line (HOL) and its being transmitted successfully, less the time used for the successful data exchange procedure. The bound can be used to predict the performance of a WLAN system.

To evaluate the effect of differentiated AIFS lengths, we set the persistence factor to be one for all virtual stations, which means the CW value will keep constant when collision happens. We justify this assumption by two results: (a) one is valid value for persistent factor according to IEEE802.11 standards [1]. (b) Persistence factor takes effect only when collision occurs, and CW is reset to CW_{\min} after a successful transmission. We argue that collision is a relatively rare occurrence during the operation of the system when the nodes use EDCF function, which will be illustrated when we discuss the numerical results. We will discuss further how this assumption affects the results in Section 5.

Our last assumption is that the network consists of finite number of nodes. Each node operates only one virtual station. In other words, we ignore the effect of internal virtual collision within a station. This assumption would not take favourable effects on throughput and delay bounds, since internal virtual collision is dealt in a way that the physical collision is avoided, and physical collision results in resource waste.

3.2. Multi-dimensional Markov chain model

As shown in Figure 1, in EDCF, the channel would be in one of three states: idle, successful transmission, and collision. The system is in the idle state when all nodes either have no packet to transmit or are executing the sensing and backoff procedure before the data transmission. The system is in the successful transmission state when one and only one node is transmitting a packet without collision. Specifically, it denotes a time interval from sending RTS frame to receiving ACK frame in RTS/CTS mode or a time interval from sending DATA frame to receiving ACK frame in basic transmission mode. The system is in collision state if more than one node tries to exchange packets at the same time. In particular, it denotes a time interval from sending RTS frame to CTS timeout at the sender in RTS/CTS mode or the time sending DATA frame to ACK timeout in basic transmission mode. Three states interleave one another on the time axis. The system throughput is the percentage of successful transmitted payload** over the total time.

Consider a fixed number (m) of flows (f_1, f_2, \dots, f_m), each of which employs one virtual station equipped with a set of TC-specific parameters, i.e. $AIFS_i$, CW_{\min_i} , CW_{\max_i} , and PF_i where $i = 1, 2, \dots, m$. The difference between any two AIFSs is non-negative integer multiple of a slot time σ .

In our derivation, a discrete and slotted time scale is adopted. Let t_n denote the time of end of n th transmission attempt of the system. As illustrated in Figure 1, $t_0, t_1, t_2 \dots$ represent such time points. The state of the stochastic process at time t_n is a vector of backoff counters of m VSs $B(n) = (b_n^{(1)}, b_n^{(2)}, \dots, b_n^{(m)})$, where $b_n^{(i)}$ is the value of backoff counter of the i th VS at time t_n . The transmission could be either a successful one or a collision, hence the duration between two consecutive end of transmission may be different. The state of the system changes as a transmission attempt occurs, and at the end of transmission attempt, new backoff counter(s) of VSs, which just completed the transmission attempt, will be randomly generated. Since the state at time t_{n+1} only relies on the state at time t_n , we can model this process as an m -dimensional discrete time Markov chain. The state space of this Markov chain is all possible combinations of $(b_n^{(1)}, b_n^{(2)}, \dots, b_n^{(m)})$, where $b_n^{(i)}$ is any integer between $[1, CW_i + 1]$, $i = 1, 2, \dots, m$.

** In this paper, the payload length is normalized by transmission rate.

For convenience and without losing generality, we use a non-negative integer to represent the AIFS length, for example, if $AIFS = k$, it means the length of AIFS is the length of DCF InterFrame Space (DIFS) plus k slot times (σ) [1]. According to the 802.11e draft, k is chosen in the range $[0, 8]$.

We next derive the one-step transition matrix P of the model. From time t_n , each node starts or resumes the carrier sensing and backoff procedure in order to initiate a packet exchange. Then some nodes will finish their backoff procedure earlier than others and proceed with data transmission. No matter whether the attempt succeeds or fails, in the end, system reaches the time point t_{n+1} . It is obvious that within the duration $[t_n, t_{n+1}]$, there is at least one VS, say VS_j , whose backoff counter reaches zero and incurs its transmission attempt, i.e. VS_j satisfies $(b_n^{(j)} + AIFS_j) = \min_{i \in (1, 2, \dots, m)} (b_n^{(i)} + AIFS_i)$. (Note there could be more than one of such j -like VS satisfying the above minimum condition.) To ease the description of the model, we denote $s, s \in (1, 2, \dots, m)$ as the number of those VSs who execute the $(n + 1)$ st transmission attempt in $[t_n, t_{n+1}]$, then $VS_{j_1}, VS_{j_2} \dots VS_{j_s}$ are those j -like VSs that have to reset their backoff counters by randomly drawing an integer between $[1, CW_{j_r} + 1]$, ($r = 1, 2, \dots, s$) at time t_{n+1} . Moreover, with the different values of s , we can distinguish the transmission status into three mutual exclusive and exhaustive groups:

- Group 1: successful transmission (when $s = 1$).
- Group 2: partial collision (when $1 < s < m$).
- Group 3: full collision (when $s = m$).

Hence the state of this Markov chain at time t_{n+1} becomes $B(n + 1) = (b_{n+1}^{(1)}, b_{n+1}^{(2)}, \dots, b_{n+1}^{(m)})$, where $b_{n+1}^{(r)}$ is any integer chosen from $[1, CW_{j_r} + 1]$ with uniform distribution, $r = 1, 2, \dots, s$.

While those VSs whose backoff counters did not count down to zero within $[t_n, t_{n+1}]$, will have their states at t_{n+1} as:

$$b_{n+1}^{(l)} (l \neq j_1, j_2, \dots, j_s) = \begin{cases} b_n^{(l)} & \text{if } AIFS_l \geq AIFS_{j_r} + b_n^{(j_r)} \\ b_n^{(l)} + AIFS_l - (AIFS_{j_r} + b_n^{(j_r)}), & AIFS_l < AIFS_{j_r} + b_n^{(j_r)} \end{cases}$$

And the one-step transition probability $P_{B(n) \rightarrow B(n+1)}$ is:

$$[(CW_{j_1} + 1) \cdot (CW_{j_2} + 1) \cdot \dots \cdot (CW_{j_s} + 1)]^{-1}$$

Figure 2 illustrates the one-step transition diagram for the collision transmission situation.

3.3. Performance analysis

The model can be regarded as a regenerative process. Starting from any full collision state, the system will eventually visit another full collision state; after that, the system will start over a probabilistic replica of the whole process, which ends with another full collision. We define the time interval between two full collision states as a round. Therefore, full collision states of this model are regenerative states and they are identical in that they restart probabilistically identical operation round. The theory of regenerative processes indicates the statistical characteristic in one round is identical to the long-run properties of the system. The following analysis and derivation is based on one round of operation.

Specifically, we define the start of each round as leaving a regenerative state, and ends in the following regenerative state. In other words, a round randomly starts from any valid Markov

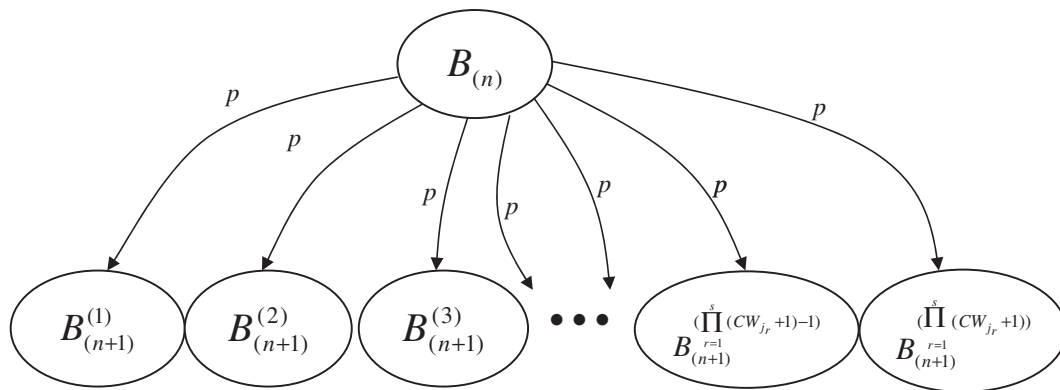


Figure 2. State transition diagram of Markov chain. $B(n)$ could transit to any of $[(CW_{j_1} + 1)(CW_{j_2} + 1) \dots (CW_{j_s} + 1)]$ possible states, each with the same probability $p = 1/[(CW_{j_1} + 1)(CW_{j_2} + 1) \dots (CW_{j_s} + 1)]$.

state after the last full collision. In order to compute the *throughput and access delay*, we adopt the theorem of *transient state and time to absorption* (refer to References [23, 24]). In one round of operation, the regenerative state, i.e. $B(n)$ with the property $b_n^{(1)} + \text{AIFS}_1 = b_n^{(2)} + \text{AIFS}_2 = \dots = b_n^{(m)} + \text{AIFS}_m$ can be taken as an absorbing state, in which case, all VSs are going to send their packets simultaneously within $[t_n, t_{n+1}]$ and reset their backoff counters at t_{n+1} consequently. The other states, including both successful and partial collision states, are transient states. Using the theorem on *time to absorption*, we can accurately calculate how many times on average the system will go through each transient state (either successful transmission or partial collision) before being absorbed into absorbing states (full collision) starting from the beginning of a round. Note that the system operation will go on with another round, instead of being absorbed in full collision states.

3.3.1. *Notations and procedure.* The computation procedure is as follows:

1. Compose one-step transition matrix P of the Markov chain; organize its layout so that the absorbing states are ordered before all transient states. Matrix P would have the form as $P = \begin{bmatrix} I & 0 \\ R & Q \end{bmatrix}$, where I is unit matrix, R consists of transition probabilities from transit states to absorbing states, and Q represents transition probabilities from transit states to transit states.
2. Calculate the absorption matrix $W = (I - Q)^{-1}$. The element of W , say W_{jk} represents the average number of times state k is visited starting from state j , where j and k are both transient states.
3. Calculate the reaching probability matrix $F = W \cdot R$. The element of F , say F_{ji} represent the probability that the system turns into absorbing state i starting from transit state j .

Denote G to be the set of transient states of the Markov chain corresponding to the matrix Q , and G^C is the set of absorbing states. The transient states of the Markov chain can be further divided into $m + 1$ subsets according to the transmission results (partial collision or successful)

and the transmitter of the successful transmission. Let $G_f, G_1, G_2, \dots, G_m$ denote those sets of states, where G_f is the set of states that results in a partial collision attempt, and G_i is the set of states that leads to a successful transmission from the VS_i , $i \in (1, 2, \dots, m)$.

3.3.2. The system throughput. We will calculate the *normalized throughput*, defined as the fraction of time when the channel is used to successfully transmit effective payload bits. To do so, we assume the system can equally likely start to operate from any one of valid states. Hence the total throughput of the system S can be expressed as,

$$S = \frac{E[PL]}{E[T_{\text{round}}]} = \frac{E[\text{successfully trans time}]}{E[\text{success trans time}] + E[\text{collision time}] + E[\text{idle time}]} \quad (1)$$

where $E[PL]$ is the average successfully transmitted payload (normalized by the channel rate) during one round, it comprises successfully transmitted payload by all VSs. Let $E[PL_i]$ denote payload transmitted by VS_i . $E[PL]$ can be expressed as:

$$E[PL] = \sum_{i \in (1, 2, \dots, m)} E[PL_i] \quad (2)$$

Since there could be more than one transient states (states belonging to set G_i) leading to a successful transmission by VS_i , so the effective payload transmitted by VS_i is the sum of transmissions from all states of G_i . W_{jk} ($j \in G, k \in G_i$) is the average time that the system stays in state k (leading to a successful transmission by VS_i) starting from state j . Here we assume one system round starts from any valid state of the whole state space S equally likely, which corresponds to the fact that backoff counter value of any VS is chosen randomly based on its own CW. Since the total number of states is $|P|$, i.e. the size of one step transition matrix, we can express the effective payload contributed by VS_i as:

$$\begin{aligned} E[PL_i] &= E[PL_i | \text{starting from absorbing states}] \cdot P\{\text{starting from absorbing states}\} \\ &+ E[PL_i | \text{starting from transient states}] \cdot P\{\text{starting from transient states}\} \\ &= 0 + \left(E[P_i] \cdot \sum_{j \in G} \sum_{k \in G_i} \frac{W_{jk}}{|W|} \right) \frac{|W|}{|P|} = \frac{1}{|P|} \cdot E[P_i] \cdot \sum_{j \in G} \sum_{k \in G_i} W_{jk} \end{aligned} \quad (3)$$

where $E[P_i]$ is the normalized average length of packets transmitted by a VS_i and $|P|$ is the size of one-step transition probability matrix P .

In (1), $E[\text{success trans time}]$ and $E[\text{collision time}]$ refer to the average channel busy time due to successful transmission and collision, respectively. As per IEEE802.11 specification, two access mechanisms are described, namely, the basic access mechanism and RTS-CTS exchange access mechanism. The two access mechanisms would have different transmission time and collision time. In our research, we will only discuss the RTS-CTS exchange mechanism, since it provides a better performance in the presence of high contention [26].

With RTS-CTS exchange, as shown in Figure 1, we can obtain the channel busy time needed for a successful transmission by the i th VS (T_s^i) and the channel busy time due to

collision (T_c) as:

$$\begin{aligned} T_s^i &= \text{RTS} + \text{SIFS} + \delta + \text{CTS} + \text{SIFS} + \delta + H + E[P_i] + \text{SIFS} + \delta + \text{ACK} + \delta \\ T_c &= \text{RTS} + \delta \end{aligned} \quad (4)$$

where $H = \text{PHY}_{\text{hdr}} + \text{MAC}_{\text{hdr}}$ is the packet header and δ is the propagation delay. Similar to the calculation of the effective payload, the time spent on successful transmission consists of all successful transmission by all VSs, which can be expressed as (we shorten success trans time as STT):

$$\begin{aligned} E[\text{STT}] &= E[\text{STT} \mid \text{starting from absorbing states}] \cdot P\{\text{starting from absorbing states}\} \\ &\quad + E[\text{STT}_i \mid \text{starting from transient states}] \cdot P\{\text{starting from transient states}\} \\ &= 0 + \sum_{i \in \{1, 2, \dots, m\}} \sum_{j \in G_i} \sum_{k \in G_i} \left(\frac{W_{jk}}{|W|} \cdot T_s^i \right) \cdot \frac{|W|}{|P|} \\ &= \frac{1}{|P|} \cdot \sum_{i \in \{1, 2, \dots, m\}} T_s^i \cdot \left(\sum_{j \in G_i} \sum_{k \in G_i} W_{jk} \right) \end{aligned} \quad (5)$$

The average channel busy time due to collisions consists of two parts. The first part is attributed to partial collisions before absorption, and its length is equal to the product of the time length of each collision and the sum of the total numbers of partial collision in each round. The second part is the time length of collision attributed to the absorbing state (a full collision). The average channel busy time can be expressed as:

$$\begin{aligned} E[\text{collision time}] &= E[\text{length of all partial collisions}] + E[\text{length of all full collision}] \\ &= E[\text{number of partial collisions}] \cdot E[\text{length of collision}] \\ &\quad + E[\text{number of full collisions}] \cdot E[\text{length of collision}] \\ &= \left(\sum_{j \in G_i} \sum_{k \in G_j} \frac{W_{jk}}{|W|} \cdot \frac{|W|}{|P|} \right) \cdot T_c + 1 \cdot T_c \\ &= \left(\frac{1}{|P|} \cdot \sum_{j \in G_i} \sum_{k \in G_j} W_{jk} + 1 \right) \cdot T_c \end{aligned} \quad (6)$$

The channel will be idle while all VSs are sensing and waiting for free channel. Therefore, all the three types of transmission: successful, partial collision and full collision contribute to the idle time.

Given any state at time t_n , $B(n) = (b_n^{(1)}, b_n^{(2)}, \dots, b_n^{(m)})$, let $j_1, j_2, \dots, j_s, 1 \leq s \leq m$ be the VSs having their backoff counters count down to zero and sending packets simultaneously within $[t_n, t_{n+1}]$. Let I_n denote the total idle time before the transmission within $[t_n, t_{n+1}]$, then I_n consists of two consecutive parts, namely, AIFS and idle time due to backoff counter counting down, so we can calculate it as $I_n = \text{AIFS}_{j_r} + b_n^{j_r}, r \in (1, 2, \dots, s)$. The average idle time can be

divided into two parts (we shorten idle time as IT):

$$\begin{aligned}
 E[IT] &= \sum_{i \in G^c} E[IT \mid \text{starting from absorbing state } i] \cdot P\{\text{starting from absorbing state } i\} \\
 &\quad + \sum_{j \in G} E[IT \mid \text{starting from transient state } j] \cdot P\{\text{starting from transient state } j\} \\
 &= \frac{1}{|P|} \cdot \sum_{l \in G^c} I_l + \left[\sum_{j \in G, k \in G} (W_{jk} \cdot I_k) + \sum_{j \in G} \sum_{l \in G^c} (F_{jl} \cdot I_l) \right] \cdot \frac{1}{|P|} \tag{7}
 \end{aligned}$$

Plugging (2), (3), (5), (6), (7) in Equation (1), we then get the system throughput as:

$$S = \frac{\sum_{i \in \{1,2,\dots,m\}} \sum_{j \in G, k \in G_i} (W_{jk} \cdot E[P_i])}{\sum_{i \in \{1,2,\dots,m\}} \sum_{j \in G, k \in G_i} (W_{jk} \cdot T_s^i) + \left(\sum_{j \in G, k \in G_f} (W_{jk} \cdot T_c) + |P| \cdot T_c \right) + \left(\sum_{j \in G, k \in G} (W_{jk} \cdot I_k) + \sum_{l \in G^c} \left(\sum_{j \in G} F_{jl} \right) I_l + \sum_{l \in G^c} I_l \right)} \tag{8}$$

where the denominator is $|P|$ multiple of the expectation of a round time, i.e. $|P| \cdot E[T_{\text{round}}]$.

3.3.3. Throughput ratios among flows. Now we can calculate the throughput ratios among the different flows. The effective payload sent by the i th VS, $i \in (1, 2, \dots, m)$, equals to the product of average successful transmission times and average packet length. So the throughput of the i th VS can be expressed by

$$S_i = \frac{E[PL_i]}{E[T_{\text{round}}]} = \frac{1/|P| \cdot E[P_i] \cdot \sum_{j \in G, k \in G_i} W_{jk}}{E[T_{\text{round}}]} \tag{9}$$

Therefore, the throughput ratios $S_1:S_2:\dots:S_m$ among VSs can be calculated as:

$$\begin{aligned}
 S_1:S_2:\dots:S_m &= \left(E[P_1] \cdot \sum_{j \in G, k \in G_1} W_{jk} \right) : \left(E[P_2] \cdot \sum_{j \in G, k \in G_2} W_{jk} \right) \\
 &\dots : \left(E[P_m] \cdot \sum_{j \in G, k \in G_m} W_{jk} \right) \tag{10}
 \end{aligned}$$

and the throughput ratios between any two VSs, u, v , can be expressed by

$$S_u:S_v = \left(E[P_u] \cdot \sum_{j \in G, k \in G_u} W_{jk} \right) : \left(E[P_v] \cdot \sum_{j \in G, k \in G_v} W_{jk} \right) \tag{11}$$

3.3.4. Access delay. We define the access delay as the time between a packet becoming HOL and the starting time of its successful transmission attempt, which is confirmed by ACK control frame (Figure 1, interval between t_0 and t_3 , for a packet in virtual station B). The access delay is also a critical measurement to evaluate QoS in wireless networks [15]. Based on our analytical model, we can compute the average access delay for a packet of individual flows. Specifically, using the concept of T_{round} , the average access delay of the packets from the i th flow, T_Q^i , is the total waiting time of the i th flow divided by the total number of the successful transmission by the same flow. The total waiting time of i th flow, which includes all the time spent for sensing

the channel idle, retransmission time due to collisions, as well as the successful transmission time by other flows, can also be interpreted as the average T_{round} less the total time for successful transmission of the i th VS within one round.

Following the derivation of (3), (4), (5) and (8), we obtain the average access delay of the i th VS, T_Q^i :

$$E[T_Q^i] = \frac{E[T_{\text{round}}] - E[T_{\text{success}}^i]}{E[N_s^i]} = \frac{|P| \cdot E[T_{\text{round}}] - T_s^i \cdot \sum_{j \in G} \sum_{k \in G_i} W_{jk}}{\sum_{j \in G} \sum_{k \in G_i} W_{jk}} \quad (12)$$

where $E[T_{\text{success}}^i]$ is the average time of a successful transmission for the i th VS, and $E[N_s^i]$ is the average number of successful transmissions by the i th VS.

4. NUMERICAL RESULTS AND ANALYTICAL MODEL VALIDATION

In order to validate our analytical model, we compare it with a simulation model. Numerical results from the analytical model are obtained using MATLAB. Our simulations are written in C++ . In the simulations, all stations operate independently in the RTS/CTS mode under the MAC protocol conforming to the specifications in Reference [1]. A summary of the constant parameter values used in both analytical model and simulation model are given in Table I (refer to Reference [27], 15.3.3 DS PHY characteristics). For simplicity, we set the same constant payload size for the packets from all flows.

We designed the first experiment scenario, which consists of two flows of different TCs. Both flows use the same CW values, i.e. $CW_{\text{HP}} = CW_{\text{LP}} = 7$. $AIFS_{\text{HP}}$ is fixed at 0, $AIFS_{\text{LP}}$ increments from 0 to 8. Table II presents a numerical comparison of throughput and throughput ratios obtained from analytical model (depicted by MAT) and simulation (depicted by SIM). Simulation results in all scenarios have a 95% confidence level with 5% confidence intervals. The comparison illustrates that difference between the two models are negligible, and that our analytical model is, indeed, accurate. Moreover, all the results reported in Figures 3–9 have also been compared with simulation results, with negligible differences observed.

Table I. Constant parameters in analytical model and simulations.

Packet payload size (bits)	8196
PHY header (bits)	192
MAC header (bits)	272
RTS (bits)	160
CTS (bits)	112
ACK (bits)	112
WM transmission rate (bits/s)	11M
Propagation delay (μs)	1
SIFS (μs)	10
Slot time σ (μs)	20
PIFS = SIFS + σ (μs)	30
DIFS = SIFS + 2σ (μs)	50

Table II. Comparison of analytical and simulation results.

AIFS _{LP} - AIFS _{HP}	S		HP throughput		LP throughput		Ratio	
	S-MAT	S-SIM	HP-MAT	HP-SIM	LP-MAT	LP-SIM	HP: LP MAT	HP : LP SIM
0	0.759	0.740	0.379	0.371	0.379	0.369	1.000	1.004
1	0.753	0.735	0.471	0.460	0.283	0.275	1.665	1.669
2	0.749	0.731	0.542	0.530	0.207	0.201	2.626	2.634
3	0.745	0.729	0.598	0.585	0.147	0.144	4.071	4.058
4	0.742	0.726	0.644	0.630	0.099	0.096	6.526	6.561
5	0.739	0.724	0.684	0.670	0.055	0.054	12.393	12.365
6	0.737	0.723	0.717	0.704	0.020	0.020	35.352	35.644
7	0.735	0.724	0.735	0.724	0.000	0.000	Inf	Inf

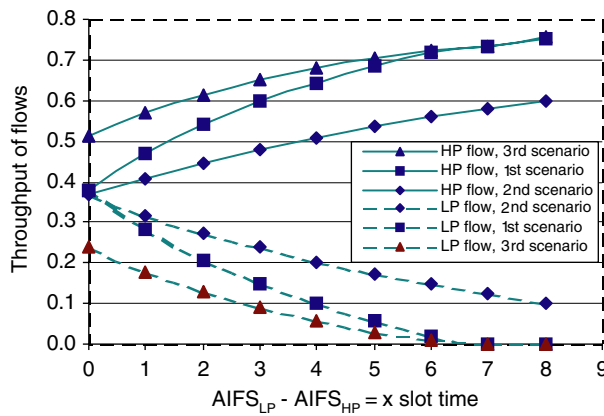


Figure 3. Throughput ratio vs AIFS difference.

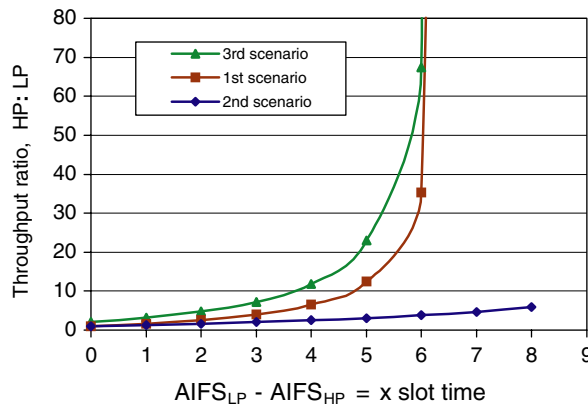


Figure 4. Throughput ratio vs AIFS difference.

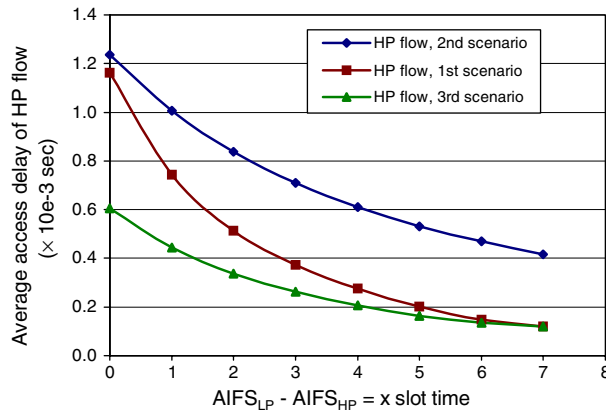


Figure 5. Access delay of HP flow vs AIFS difference.

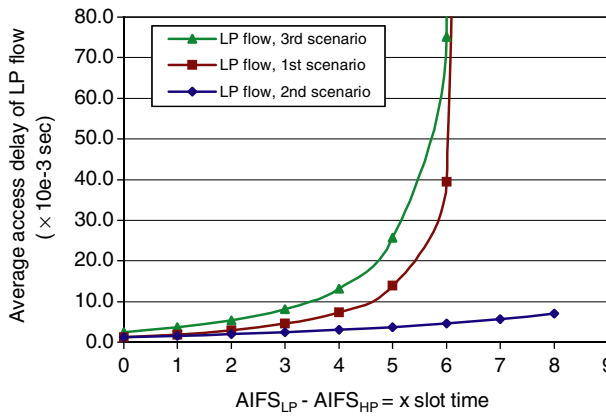


Figure 6. Access delay of LP flow vs AIFS difference.

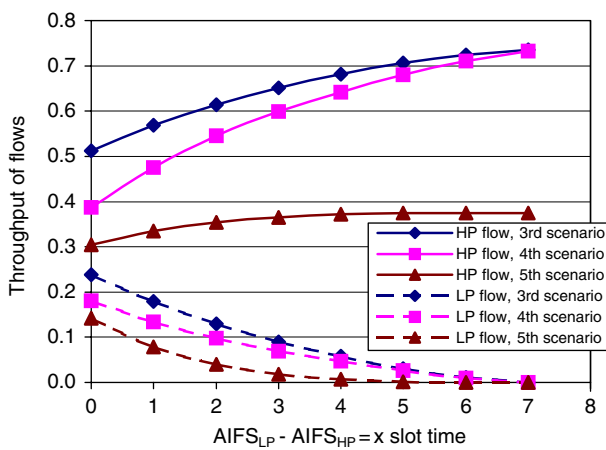


Figure 7. Impact of traffic load on flow throughput ratio.

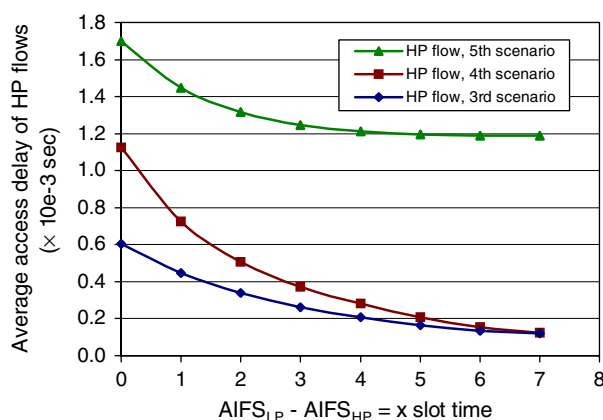


Figure 8. Impact of traffic load on access delay of HP flow.

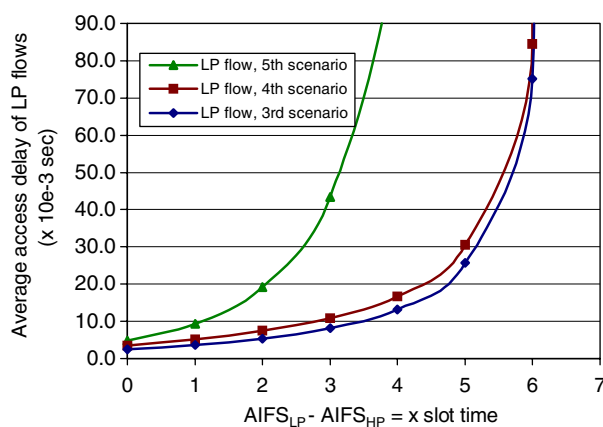


Figure 9. Impact of traffic load on access delay of LP flow.

5. PERFORMANCE EVALUATION

In this section, we investigate the performance of differentiated QoS supported by the 802.11e EDCF MAC scheme. The metrics include throughput of the system, throughput and access delay of individual flows, and throughput ratio among flows. The throughput and access delay of individual flows indicates how the flows are served distinctively as a result of TC-specific prioritized parameters, while the throughput ratio reflects quantitatively the QoS differentiation, i.e. to what degree, the HP TCs have an advantage over LP ones.

5.1. The compound effect of AIFS And CW on QoS

To explore the comprehensive impact of differentiated AIFSs and CWs, we design another two sets of experiments. The second scenario consists of two flows and each flow has the same

parameters as the above experiments, except that both flows use $CW = 15$; and the LP flow increases AIFS length by one slot time each time. The third scenario also consists of two flows. However, the HP flow has $CW_{HP} = 7$, the LP flow has $CW_{LP} = 15$; i.e. both AIFSs and CWs are differentiated.

Figure 3 illustrates throughput of the HP flow and the LP flow of the three sets of experiments. Figure 4 shows the throughput ratios between the HP and LP flows in each scenario. Figures 5 and 6, respectively, show access delay of the HP flow and LP flow of the experiments.

We make the following observations. First, we note that the smaller the CWs are, the more significant the influence of AIFS on QoS differentiation will be. Secondly, the combination of differentiated AIFSs and differentiated CWs introduce more compound and significant effect on the degree of differentiation. Obviously, scenario three shows a more intensive differentiation in term of both throughput and access delay with the increase of AIFS difference. Moreover, we argue that it is the AIFS difference, rather than the absolute AIFS values, that determines the degree of QoS differentiation. As displayed in Figure 4, the increment of throughput ratio accelerates with the growth of AIFSs difference non-linearly. The larger the AIFSs difference, the faster the ratio increases, and in the worst case, the LP flow may completely lose the opportunity to access medium.

5.2. Effect of traffic loading on QoS

Here we evaluate the performance as the traffic load changes. Figures 7, 8 and 9, respectively, show the results of the throughput of HP flow and LP flow, access delay of HP flow, and access delay of LP flow. Each HP flow adopts the TC-specific parameters as $CW_{HP} = 7$ and $AIFS_{HP} = 0$; Each LP flow has TC-specific parameters as $CW_{LP} = 15$ and $AIFS_{LP}$ is incremented from 0 to 7 as the experiment proceeds. Each figure plots the following cases: (a) scenario three: one HP flow and one LP, (b) scenario four: one HP flow and two identical LP flows (c) scenario five: two identical HP flows and one LP flow. From the results we can see that the HP flows can dominate the share of wireless medium and are less affected by LP flows. To our surprise, a small difference in AIFS (one slot time) between two flows can result in considerable large difference of their throughput and access delay when there is one more HP flow.

5.3. Discussion

In order to advocate the argument that collisions are relatively infrequent occurrences compared to successful transmissions, we further investigate the experimental results based on the analytical model. Table III summarizes the results of average number of collisions and successful transmissions during one round of operation, in two cases.

Table III. Summary of collisions and transmission attempts.

	Case one—one HP and two LP	Case two—two HP and one LP
Transmission attempts in one round	194.9	196.9
Collisions in one round	16.8	27.3
Ratio of collisions over total transmission	0.094	0.139

The second case is a snapshot of the fourth scenario, consists of one HP flow and two LP flows. The HP flow and the LP flows use the same parameters as above. The ratio of collision over the total transmission attempts is 0.094.

The first case is a snapshot of the fifth scenario, which composes of two HP flows and one LP flow. Both HP flows use the parameters as $AIFS_{HP} = 0$, and $CW_{HP} = 7$; while LP flow uses the parameters as $AIFS_{LP} = 3$ and $CW_{LP} = 15$. The ratio of collision over the total transmission attempts is 0.139.

We set PF to one for all experiment—an extreme case. However, since the collisions are infrequent during the operation of system, our results should be representative for other valid PF value cases. Furthermore, since a PF value greater than one might help to reduce the happening of collisions, so the throughput results using PF as one would be the upper bound value in the corresponding experimental cases.

6. CONCLUDING REMARKS

In this paper, we developed an accurate multi-dimensional Markov model to analyse the performance of the IEEE 802.11e EDCF MAC protocol. Particularly, the model can be used to evaluate the ability of the IEEE 802.11e EDCF to provide differentiated QoS. Based on the proposed analytical model, we have derived formulations of saturation throughput, throughput ratio of flows and access delay as functions of the prioritized parameters. It also has been used to conduct quantitative analysis of the impact of parameters such as AIFS and CW on the performance of prioritized flows.

Numerical results show that EDCF can well support differentiated QoS, and it provides significant advantage to higher priority flows. Numerical results show that AIFS has a significant impact on the TC priority. Fixing other parameters, the smaller the AIFS is, the higher the priority of a flow, and the shorter time the flow will wait before transmitting. This in turn translates into higher bandwidth share of higher priority flows. Hence, we conclude that the IEEE 802.11 EDCF function can effectively provide QoS differentiation.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees for their useful comments on an earlier version of this paper. Their suggestions helped improve the quality of the paper. This work has been supported by grants from Communications and Information Technology Ontario (CITO) and the Natural Science and Engineering Research Council of Canada (NSERC). This support is greatly appreciated.

REFERENCES

1. IEEE 802.11 WG. *Draft Supplement to Standard for Telecommunications and Information Exchange Between Systems—LAN/MAN Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancement for Quality of Service (QoS)*. 2001.
2. Bharghavan V. MACAW: a media access protocol for wireless LAN's. *Sigcomm94*, London, U.K., September 1994.
3. Qiao D, Shin KG. Achieving efficient channel utilization and weighted fairness for data communications in IEEE 802.11 WLAN under the DCF. *Proceedings of the Tenth International Workshop on Quality of Service (IWQoS'2002)*, Miami Beach, FL, 15–17 May 2002.

4. Bensaou B, Yu Wang, Chi Chung Ko. Fair medium access in 802.11 based wireless *ad-hoc* networks. *First Annual IEEE and ACM International Workshop on Mobile Ad Hoc Networking and Computing (MobiHoc)*, Boston, MA, U.S.A., August 2000.
5. Timucin Ozugur. Optimal MAC-layer fairness in 802.11 networks. *ICC'02*, 2002.
6. Timucin Ozugur. Weighted hierarchical backoff algorithm for wireless *ad hoc* networks. *Globecom'01*, 2001.
7. Nandagopal T, Kim T-E, Gao X, Bharghavan V. Achieving MAC layer fairness in wireless packet networks. *Proceedings of the MOBICOM*, Boston, 2000.
8. Barry M, Campbell AT, Veres A. Distributed control algorithms for service differentiation in wireless packet networks. *INFOCOM 2001*, vol. 1, 2001; 582–590.
9. Veres A, Campbell AT, Barry M, Li-Hsiang Sun. Supporting service differentiation in wireless packet networks using distributed control. *IEEE Journal on Selected Areas in Communications* 2001; **19**(3):2081–2093.
10. Hsiao-Kuang Wu, Pei-Hung Chuang. Dynamic QoS allocation for multimedia *ad hoc* wireless networks. *Mobile Networks and Application* 2001; **6**:377–384.
11. Aad I, Castelluccia C. Differentiation mechanisms for IEEE802.11. *IEEE Infocom 2001*, Anchorage, AK, 2001.
12. Seung-Seok Kang, Mutka MW. Provisioning service differentiation in *ad hoc* networks by modification of the backoff algorithm. *International Conference on Computer Communication and Network (ICCCN) 2001*, October 2001.
13. Vaidya NH, Bahl P, Gupta S. Distributed fair scheduling in a wireless LAN. *Proceedings of MOBICOM*, Boston, MA, August 2000.
14. Banchs A, Perez X. Distributed weighted fair queuing in 802.11 wireless LAN. *ICC2001*, Helsinki, Finland, 11–15 June 2001.
15. Ahn G-S, Campbell AT, Veres A, Li-Hsiang Sun. Supporting service differentiation for real-time and best effort traffic in stateless wireless *ad hoc* networks (SWAN). *IEEE Transactions on Mobile Computing* 2002; **1**(3):182–207.
16. Mangold S, Choi S, May P, Klein O, Hiertz G, Stibor L. IEEE802.11e wireless LAN for quality of service. *Proceedings of the European Wireless*, vol. 1, Florence, Italy, February 2002; 32–39.
17. Grilo A, Nunes M. Performance evaluation of IEEE802.11e. *PIMRC 2002*, Portugal, 15–18 September 2002.
18. Lindgren A, Almquist A, Schelen O. Quality of service schemes for IEEE 802.11 wireless LANs: an evaluation. *Mobile Networks and Applications* 2003; **8**(3):223–235.
19. Romdhani L, Ni Q, Turletti T. AEDCF: enhanced service differentiation for IEEE 802.11 wireless *ad-hoc* networks. *INRIA Research Report No. 4544*, 2002.
20. Romdhani L, Ni Q, Turletti T. Adaptive EDCF: enhanced service differentiation for IEEE802.11 wireless *ad hoc* networks. *IEEE WCNC'03*, New Orleans, U.S.A., 2003.
21. Ksentini A, Naimi M, Nafaa A, Gueroui M. Adaptive service differentiation for QoS provisioning in IEEE 802.11 wireless *ad hoc* networks. *Proceedings of the 1st ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks* 2004; 39–45.
22. Banchs A, Perez-Costa X, Qiao D. Providing throughput guarantees in IEEE 802.11e wireless LANs. *The 18th International Teletraffic Congress*, Berlin, Germany, 2003.
23. Ross SM. *Introduction to Probability Models* (7th edn). Academic Press: New York, 2000.
24. Kao EPC. *An Introduction to Stochastic Processes*. Duxbury Press: Belmont, CA, U.S.A., 1997.
25. Rappaport TS. *Wireless Communications—Principles and Practice*. Prentice-Hall: Englewood Cliffs, NJ, 1996.
26. Bianchi G. Performance analysis of the IEEE802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications* 2000; **18**(3):235–547.
27. IEEE 802.11 WG. *Reference number ISO/IEC 8802-11:1999(E) IEEE Std 802.11, International Standard [for] Information Technology—Telecommunications and Information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. 1999.

AUTHORS' BIOGRAPHIES



Kenan Xu received the BEng degree in computer engineering from University of Electronics Science and Technology of China (UESTC) in 1997, and the MAsc degree from Queen's University, Kingston, Canada, in 2003. He is currently pursuing the PhD degree in computer engineering at Queen's university. He is the primary author and/or co-author of a number of journal, and international technical conference papers as well as two book chapters. He acted as reviewer for several technical journals and international conferences, including *Computer Communications Journal*, *IEEE ISCC 2003*, *IEEE ISCC 2004*, *IEEE IPCCC 2004*, *Networking 2004*, *IEEE GLOBECOM 2004* and *IEEE ICC 2005*. From 1997 to 2000, he worked for TJNEC corporations as a software engineer, In 2001, he worked with Telos Technology, Richmond, Canada, as a software engineer. During his study at

Queen's, he served as a research assistant in the Telecommunication Research Lab (TRL) and a teaching assistant in the Department of Electrical & Computer Engineering. In 1997, he was awarded the excellent graduate award as the first class graduate student of bachelor degree. In 2004, he received an NSERC PhD scholarship. His current research interests include wireless networks, QoS architectures, protocols and algorithms, energy-efficient schemes, performance analysis and evaluation, and wireless sensor networks.



Quanhong Wang received her BSc and MSc degrees in Electrical Engineering from Harbin Institute of Technology and Beijing University of Aeronautics and Astronautics, China, respectively. She is currently working towards her PhD degree at Department of Electrical and Computer Engineering of Queen's University, Canada. She is a student member of the IEEE and the IEEE Communications Society. She has authored several papers in top-tier journals and conferences in the area of wireless communications and networks. Since 2002, she has been a research assistant of the Telecommunications Research Lab (TRL) at Queen's University and a teaching assistant in the Department of Electrical & Computer Engineering. In 2003 and 2004, she has been awarded Ontario Graduate Scholarship for two consecutive years. Her current research interests are in the area of Wireless Ad Hoc and sensor Networks with special emphasis on energy efficient architecture design, topology control, protocols and algorithms and performance evaluation.



Hossam Hassanein is a leading researcher in the School of Computing at Queen's University in the areas of broadband, wireless and variable topology networks architecture, protocols, control and performance evaluation. Before joining Queen's University in 1999, he worked at the Department of Mathematics and Computer Science at Kuwait University (1993–1999) and the Department of Electrical and Computer Engineering at the University of Waterloo (1991–1993). Dr Hassanein obtained his PhD in Computing Science from the University of Alberta in 1990. He is the founder and director of the Telecommunication Research (TR) Lab <http://www.cs.queensu.ca/~trl> in the School of Computing at Queen's. Dr Hassanein has more than 150 publications in reputable journals, conferences and workshops in the areas of computer networks and performance evaluation. Dr Hassanein has served on the program committee of a number international conferences and workshops.

He is the editor of the IEEE 802.6e standard for Slot Reuse in Distributed Queue Dual Bus networks in 1992. Dr Hassanein is the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research award in 2003. He is also the recipient of the 1993 IEE 'Hartee Premium' best paper award, for the IEE Proceedings on Computers.