Supporting QoS in IEEE 802.11e Wireless LANs

Xiang Chen, *Member, IEEE,* Hongqiang Zhai, *Student Member, IEEE,* Xuejun Tian, and Yuguang Fang, *Senior Member, IEEE*

***Abstract*— In the emerging IEEE 802.11e MAC protocol, the Enhanced Distributed Channel Access (EDCA) is proposed to support prioritized QoS; however, it cannot guarantee strict QoS required by real-time services such as voice and video without proper network control mechanisms. To overcome this deficiency, we first build an analytical model to derive an average delay estimate for the traffic of different priorities in the unsaturated 802.11e WLAN, showing that the QoS requirements of the real-time traffic can be satisfied if the input traffic is properly regulated. Then, we propose two effective call admission control schemes and a rate control scheme that relies on the average delay estimates and the channel busyness ratio, an index that can accurately represent the network status. The key idea is, when accepting a new real-time flow, the admission control algorithm considers its effect on the channel utilization and the delay experienced by existing real-time flows, ensuring that the channel is not overloaded and the delay requirements are not violated. At the same time, the rate control algorithm allows the best effort traffic to fully use the residual bandwidth left by the real-time traffic, thereby achieving high channel utilization.**

***Index Terms*— QoS, 802.11e, Wireless LAN.**

I. Introduction

WIRELESS LANs based on the IEEE 802.11 Distributed Coordination Function (DCF) [13] have been widely used in recent years due to their simple deployment and low cost. Since the current DCF can only support best effort traffic, the IEEE 802.11 Task Group E recently proposed a new contention-based channel access method called Enhanced Distributed Channel Access (EDCA) in the IEEE 802.11e standard [14] [6] [20]. Despite providing prioritized quality of service (QoS), the EDCA still cannot support strict QoS for real-time applications like voice and video [6]. This paper studies how the EDCA can be enhanced to meet this challenge.

The creation of the EDCA is due to extensive research works that aimed to support prioritized service over the 802.11 DCF [1] [23] [24] [27]. Ada and Castelluccia [1] proposed to scale the contention window and use different inter frame spacing or maximum frame length for services of different priorities. In [24], two mechanisms, i.e., virtual MAC and virtual source, were proposed to enable each node to provide differentiated services for voice, video, and data. To further protect the real-time traffic in the 802.11e WLAN, Xiao et al. adopted a two-level mechanism [27]. In summary, although these works succeed in supporting service differentiation, strict QoS cannot be satisfactorily addressed.

Meanwhile, considerable effort was devoted to theoretical analysis of the performance of the 802.11 DCF [3] [5] [8] [12] [25] [26] [28]. In [3], Bianchi proposed a Markov chain model for the binary exponential backoff procedure. By assuming the collision probability of each node's transmission is constant and independent of the number of retransmissions, he derived the saturated throughput for the IEEE 802.11 DCF. Based on the saturated throughput derived in Bianchi's model, Foh and Zuckerman [8] used a Markovian state dependent single server queue to analyze the throughput and mean packet delay. Cali et al. [5] studied the 802.11 protocol capacity by using a p-persistent backoff strategy to approximate the original backoff in the protocol. Recently, we derived an approximate probability distribution of the service time, and based on the distribution, analyzed the throughput and average service time [28]. In [26], Xiao proposed an analytical model to evaluate the performance of the 802.11e in the saturated case. Clearly, no analysis were focused on the performance of the EDCA in the unsaturated case.

In our previous work [30], we have found that it is in the unsaturated case that the 802.11 achieves the maximum throughput and small delay because of the low collision probability; by contrast, when working in the saturated case, it suffers from a large collision probability, leading to low throughput and excessively long delay. Motivated by this discovery, we aim to tune the network to work in the unsaturated case in order to support strict delay requirements of real-time services. However, effective tuning is not easy to achieve given that the 802.11 EDCA is in nature contention-based and distributed, thereby making it hard to characterize actual traffic conditions in the network. To overcome these difficulties, we propose two call admission control schemes and a rate control scheme that function based on the novel use of the channel busyness ratio. It is important to note that while the IEEE 802.11e recommends the use of call admission control, no algorithm is specified. In addition, the IEEE 802.11e has not addressed any issue on rate control.

In this paper, we make the following contributions. First, we build an analytical model to derive an average delay estimate for the traffic of different priorities in the unsaturated 802.11e

Manuscript received November 2004; revised July 21, 2005 and September 16, 2005; accepted September 16, 2005. The associate editor coordinating the review of this paper and approving it for publication was J. Zhang. The work of X. Chen, H. Zhai, and Y. Fang was supported in part by the US Office of Naval Research, under grant N000140210464 (Young Investigator Award) and under grant N000140210554, and by the US National Science Foundation, under grant ANI-0093241 (CAREER Award) and under grant ANI-0220287. The work of X. Tian was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Exploratory Research, 17650020, 2005.

X. Chen is with Motorola Labs (e-mail: a00131@motorola.com).

H. Zhai and Y. Fang are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 (e-mail: zhai@ecel.ufl.edu, fang@ece.ufl.edu).

X. Tian is with the Department of Information Systems, the Faculty of Information Science and Technology, Aichi Prefectural University, Aichi, Japan (e-mail: tan@ist.aichi-pu.ac.jp).

Digital Object Identifier 10.1109/TWC.2006.04762.

wireless LAN. We show that if the traffic is properly regulated, the 802.11e WLAN is capable of supporting QoS requirements for the real-time traffic. The delay estimate is then used in the call admission control. Second, since the channel busyness ratio is easy to obtain and can accurately represent the network status, it provides a very suitable control variable for both the call admission control and the rate control. As a result, the call admission and rate control schemes are simple and effective. The admission control over the real-time traffic guarantees its QoS requirements can be satisfied, and the rate control allows the best effort traffic to make full use of the residual channel capacity while not affecting QoS of the real-time traffic.

The remainder of this paper is organized as follows. In Section II, we give a brief introduction of the IEEE 802.11e EDCA. The delay performance is analyzed in Section III, and verified in Section IV. We then present our proposed schemes in Section V. In Section VI, the performance is evaluated through comprehensive simulation studies. Finally, Section VII concludes this paper.

II. Operations of the IEEE 802.11E

The legacy IEEE 802.11 DCF (Distributed Coordination Function) is based on carrier sense multiple access with collision avoidance (CSMA/CA). Before starting a transmission, each node performs a backoff procedure, with the backoff timer uniformly chosen from [0, CW-1] in terms of time slots, where CW is the current contention window. If the channel is determined to be idle for a backoff slot, the backoff timer is decreased by one. Otherwise, it is suspended. When the backoff timer reaches zero, the node transmits a DATA packet. If the receiver successfully receives the packet, it acknowledges the packet by sending an acknowledgment (ACK) after an interval called short inter-frame space (SIFS). So this is a two-way DATA/ACK handshake. If no acknowledgment is received within a specified period, the packet is considered lost; so the transmitter will double the size of CW and choose a new backoff timer, and start the above process again. When the transmission of a packet fails for a maximum number of times, the packet is dropped. To reduce collisions caused by hidden terminals and improve channel efficiency for long data transmissions [2], the RTS/CTS (request to send/clear to send) mechanism is employed. Therefore, a four-way RTS/CTS/DATA/ACK handshake is used for a packet transmission.

Based on the DCF, the EDCA is meant to provide prioritized services. In the EDCA, traffic of different priorities is assigned to one of four transmit queues, which respectively correspond to four access categories (ACs). Each queue transmits packets with an independent channel access function, which implements the prioritized channel contention algorithm. In other words, different channel access functions use different contention windows (the minimum and maximum contention windows) and backoff timers. Specifically, for AC i ($i = 0, 1, 2, 3$), the initial backoff window size is $CW_{min}[i]$, the maximum backoff window size is $CW_{max}[i]$, and the arbitration inter-frame space is $AIFS[i]$. For $0 \leqslant i < j \leqslant 3$, $CW_{min}[i] \geqslant CW_{min}[j]$, $CW_{max}[i] \geqslant CW_{max}[j]$, and $AIFS[i] \geqslant AIFS[j]$. Note that in the above inequalities, at least one must be strictly "not equal to". Thus, we see that the AC with a higher level has a higher priority, since it has a higher probability to gain channel access. When an application is admitted, it will be attached with a specific priority and assigned to the corresponding queue, which performs like a single node in the DCF.

III. Delay Analysis of the IEEE 802.11E

This section focuses on the delay analysis of the IEEE 802.11e EDCA in the unsaturated case, where the collision probability is small and packets do not accumulate in the transmit queue. We consider the case where the RTS/CTS mechanism is used, although our analysis can be extended to the basic access mechanism. The channel is assumed to be perfect, i.e., no packet is lost due to channel fading. In accordance with the IEEE 802.11e protocol, there are at most four transmit queues in each active nodes. Let i (= 0, 1, 2, 3) denote the priority of the four queues, with $i = 3$ being the highest priority. Also, let n_i denote the number of queues of priority i in the network. Each priority queue is treated as an independent node. Next, we first distinguish between the saturated case and the unsaturated case.

A. Saturated Case vs. Unsaturated Case

By saturation, we mean the network is overloaded and each node always has packets to transmit. In other words, the transmit queue at each node is always not empty. As a matter of fact, all the nodes will keep contending for the channel, leading to a high level of packet collisions especially in the presence of a large number of nodes. As a result, the packet cannot get through and the transmit queue will build up and cause packet losses due to buffer overflow. On the contrary, if the network works in the unsaturated case, not all the nodes are contending for the channel at the same time. Therefore, the packet collision is low and packets get transmitted quickly. Also, the queue is not always nonempty. In this case, we need to explicitly consider this possibility when building the analytical model.

While the saturated throughput was shown be stable when the network is overloaded [3], we have demonstrated that the maximum throughput is achieved in the unsaturated case and the difference becomes more visible when the number of active nodes is fairly large [30]. Furthermore, in the saturated case, the packet collision probability given the number of nodes in the network is the highest, leading to long MAC service time. Also, the queue build-up results in long queueing delay. Clearly, the saturated case is undesirable to support real-time traffic that has strict delay requirement.

B. Markov Chain Model for the IEEE 802.11e

Consider a priority i queue. We define $b(i, t)$ as a stochastic process representing the value of the backoff counter at time t, and $s(i, t)$ as a stochastic process representing the backoff stage j, where $0 \leqslant j \leqslant \alpha$. Here α is the maximum number of retransmissions and is equal to 7 according to the standard. Let $CW_{i,min}$ and $CW_{i,max}$ be the minimum and maximum contention window for priority i, then $CW_{i,max} = 2^m CW_{i,min}$, where m is the maximum number of the stages allowed in the

exponential backoff procedure and is equal to 5 according to the standard. For convenience, we define $W_{i,0} = CW_{i,min}$. Therefore, at different backoff stage $j \in (0, \alpha)$, the contention window size

$$W_{i,j} = \begin{cases} 2^j W_{i,0} & if\ 0 \leqslant j \leqslant m \\ 2^m W_{i,0} & if\ m < j \leqslant \alpha. \end{cases} \tag{1}$$

Let p_i denote the probability of collision seen by a transmitted packet from queue i. Similar to [3] [30], if p_i is assumed to be independent of the backoff procedure, then the two-dimensional process $\{s(i,t), b(i,t)\}$ can be modeled as a discrete-time Markov chain, as shown in Fig. 1, where the state couplet $(j,\ k)$ means that the backoff stage is j and the backoff counter is k. In this Markov chain, the only non-null one step transition probabilities are as follows.

$$\begin{cases} P\{j,k|j,k+1\} = 1 \quad k \in (0, W_{i,j}-2) \quad j \in (0,\alpha) \\ P\{0,k|j,0\} = (1-p_i)/W_{i,0} \quad k \in (0, W_{i,0}-1) \quad j \in (0,\alpha-1) \\ P\{j,k|j-1,0\} = p_i/W_{i,j} \quad k \in (0, W_{i,j}-1) \quad j \in (1,\alpha) \\ P\{0,k|\alpha,0\} = 1/W_{i,0} \quad k \in (0, W_{i,0}-1), \end{cases} \tag{2}$$

where $P\{j_1,k_1|j_0,k_0\} = P\{s(i,(t+1)) = j_1, b(i,(t+1)) = k_1 | s(i,t) = j_0, b(i,t) = k_0)\}$.

Letting $b_{j,k} = \lim_{t\to\infty} P\{s(i,t) = j, b(i,t) = k\}$ be the stationary distribution of the chain, we have

$$b_{j-1,0} p_i = b_{j,0} \quad 0 < j \leqslant \alpha. \tag{3}$$

from which we obtain

$$b_{j,0} = p_i^j b_{0,0} \quad 0 \leqslant j \leqslant \alpha. \tag{4}$$

Because of the chain regularities, for each $k \in (1, W_{i,j}-1)$, $b_{i,k}$ can be expressed as

$$b_{j,k} = \frac{W_{i,j}-k}{W_{i,j}} \times \begin{cases} (1-p_i)\sum_{l=0}^{\alpha-1} b_{l,0} + b_{\alpha,0} & j = 0 \\ p_i b_{j-1,0} & 0 < j \leqslant \alpha. \end{cases} \tag{5}$$

Given Equation (4), Equation (5) can be simplified as

$$b_{j,k} = \frac{W_{i,j}-k}{W_{i,j}} b_{j,0} \quad 0 \leqslant j \leqslant \alpha. \tag{6}$$

Furthermore, by using the normalization condition

$$1 = \sum_{j=0}^{\alpha} \sum_{k=0}^{W_{i,j}-1} b_{j,k} = \sum_{j=0}^{\alpha} b_{j,0} \sum_{k=0}^{W_{i,j}-1} \frac{W_{i,j}-k}{W_{i,j}} = \sum_{j=0}^{\alpha} b_{j,0} \frac{W_{i,j}-1}{2}, \tag{7}$$

we obtain Equation (8).

Therefore, the probability that a node of priority i transmits in a random slot, given that the queue is not empty, denoted by τ_i, is obtained in Equation (9).

Once τ_i is known, p_i can be obtained as in Equation (10), where $P_{i,0}$ is the probability that a priority i queue is empty.

C. G/M/1 Queue Model to Estimate Mean Delay

We model a priority i queue as a queueing system in order to derive the probability $P_{i,0}$. In the queueing system, the packet arrival process is determined by the traffic characteristics of a priority i application that emits packets to the MAC layer. Without loss of generality, we assume the packet interarrival time is generally distributed. The service time of the queueing system, which is also called the *MAC layer service time*, is the time period from the instant that a packet moves to the head of the queue and begins to be serviced by the MAC layer to the instant that it is successfully transmitted or dropped after all the α times of retransmissions fail. As shown in our prior work [28], we have derived the probability generating function (PGF) of the MAC service time and hence its probability distribution. We also demonstrated that the MAC layer service time can be well approximated with the exponential distribution. Thus, for a priority i queue, the service time is exponentially distributed with mean $1/\mu_i$, where $1/\mu_i$ can be obtained from the PGF and expressed as a function of the collision probability p_i, τ_i, and $P_{i,0}$. The queueing system now can be characterized by a G/M/1 queueing model.

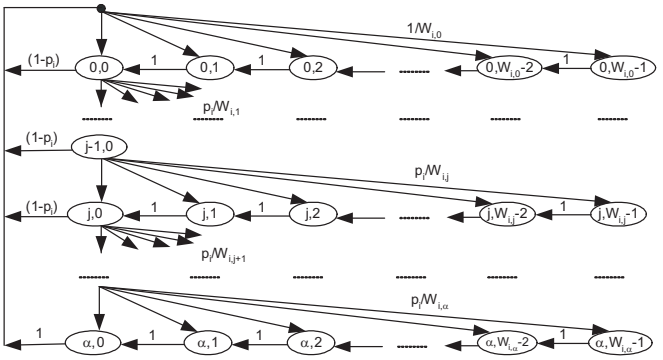


Fig. 1. Markov chain for the 802.11e backoff procedure.

Meanwhile, the probability $P_{i,0}$ can be obtained as

$$P_{i,0} = 1 - \frac{\lambda_i}{\mu_i}, \tag{11}$$

where λ_i is the average packet arrival rate for priority i traffic and is known in the traffic specification. Thus, given n_i (i = 0, 1, 2, 3) is known, we can use numerical methods to solve the nonlinear system represented by Equations (9)(10)(11) and obtain the unknown parameters p_i, τ_i, and $P_{i,0}$. Note that all these parameters lie in the interval $(0, 1)$.

Once these parameters become known, μ_i (i = 0, 1, 2, 3) is also solved. Now we can obtain the average delay experienced by a packet of priority i. In the G/M/1 system, if the probability distribution function (PDF) of the packet interarrival time is denoted by $A(t)$ and the corresponding Laplace transform is denoted by $A^*(s)$, the average system time, i.e., the average packet delay T_i, that a packet experiences can be expressed as [18]

$$T_i = \frac{1}{\mu_i(1-\sigma_i)}, \tag{12}$$

where σ_i is the unique root of

$$\sigma_i = A^*(\mu_i - \mu_i\sigma_i) \tag{13}$$

in the range of $0 < \sigma_i < 1$.

Two important points are noted. First, as seen from the above equations, we know the average delay can be obtained as long as the arrival process is of a rational Laplace transform. Clearly, this is true for most distributions. In particular, this is true for CBR traffic that has a deterministic interarrival distribution, and for VBR traffic that can be modeled with an on/off traffic model [11] [7]. Second, in the G/M/1 system,

$$b_{0,0} = \begin{cases} \frac{2(1-2p_i)(1-p_i)}{W_{i,0}(1-(2p_i)^{\alpha+1})(1-p_i)+(1-2p_i)(1-p_i^{\alpha+1})} & \alpha \leqslant m \\ \frac{2(1-2p_i)(1-p_i)}{W_{i,0}(1-(2p_i)^{m+1})(1-p_i)+(1-2p_i)(1-p_i^{\alpha+1})+W_{i,0}2^m p_i^{m+1}(1-2p_i)(1-p_i^{\alpha-m})} & \alpha > m. \end{cases} \tag{8}$$

$$\tau_i = \sum_{j=0}^{\alpha} b_{j,0} = \begin{cases} \frac{2(1-2p_i)(1-p_i^{\alpha+1})}{W_{i,0}(1-(2p_i)^{\alpha+1})(1-p_i)+(1-2p_i)(1-p_i^{\alpha+1})} & \alpha \leqslant m \\ \frac{2(1-2p_i)(1-p_i^{\alpha+1})}{W_{i,0}(1-(2p_i)^{m+1})(1-p_i)+(1-2p_i)(1-p_i^{\alpha+1})+W_{i,0}2^m p_i^{m+1}(1-2p_i)(1-p_i^{\alpha-m})} & \alpha > m. \end{cases} \tag{9}$$

$$p_i = 1 - \prod_{l=0}^{i-1}(1-(1-P_{l,0})\tau_l)^{n_l}(1-(1-P_{i,0})\tau_i)^{n_i-1}\prod_{l=i+1}^{3}(1-(1-P_{l,0})\tau_l)^{n_l}, \tag{10}$$

the buffer size is assumed to be infinite. In fact, since we focus on the unsaturated case, where p_i is small, the number of packets waiting in the queue is also small, as shown in [30]. Therefore, the above analysis is almost independent of the actual buffer size.

D. G/G/1 Queue Model to Estimate Mean Delay

More generally, we can model the service time with a general distribution with mean $1/\mu_i$ and variance $\sigma_{B_i}^2$, both of which can be obtained from the PGF and expressed as a function of the collision probability p_i, τ_i, and $P_{i,0}$. Accordingly, a priority i queue is modeled as a G/G/1 system.

Since Equation (11) still holds for a G/G/1 system, again by solving Equations (9)(10)(11) together, we obtain unknown parameters p_i, τ_i, and $P_{i,0}$, and subsequently obtain $1/\mu_i$ and $\sigma_{B_i}^2$. Now we can approximate the average delay as follows. It is well known that there exists an upper bound for the average waiting time in the queue [19]

$$W_i \leqslant \frac{\lambda(\sigma_{A_i}^2+\sigma_{B_i}^2)}{2(1-\rho_i)}, \tag{14}$$

where W_i is the average waiting time, $\rho_i = \lambda_i/\mu_i$ is the traffic intensity, and $\sigma_{A_i}^2$ and $\sigma_{B_i}^2$ are, respectively, the variances of the interarrival time and service time. This bound gets better as $\rho_i \to 1$; however, this is not the case for the unsaturated case where ρ_i is relatively small and no queue builds up. Hence, we use a weighting factor, $\frac{(\rho_i^2\sigma_{A_i}^2+\sigma_{B_i}^2)}{(\sigma_{A_i}^2+\sigma_{B_i}^2)}$, to scale down the bound to achieve a good estimate [10]. Then, the average waiting time for a packet in the queue can be approximated as

$$\widehat{W_i} = \frac{\lambda_i(\sigma_{A_i}^2+\sigma_{B_i}^2)}{2(1-\rho_i)} \times \frac{(\rho_i^2\sigma_{A_i}^2+\sigma_{B_i}^2)}{(\sigma_{A_i}^2+\sigma_{B_i}^2)} = \frac{\rho_i(\lambda_i^2\sigma_{A_i}^2+\mu_i^2\sigma_{B_i}^2)}{2\mu_i(1-\rho_i)}. \tag{15}$$

Then, the average packet delay that a packet experiences is the sum of the average waiting time in the queue and the average MAC service time $1/\mu_i$:

$$T_i = \frac{\lambda_i(\rho_i^2\sigma_{A_i}^2+\sigma_{B_i}^2)}{2(1-\rho_i)} + 1/\mu_i. \tag{16}$$

IV. Model Validation

In this section, we validate our analytical results through simulations. We consider two kinds of real-time traffic, i.e., VBR voice traffic and CBR video traffic. It is known that the interarrival time for CBR traffic is deterministic. However, it is more complicated for VBR traffic that can be modeled with an on/off traffic model [11] [7]. In this model, active periods, which are exponentially distributed with mean T_{on}, alternate with idle periods, which are exponentially distributed with mean T_{off}, according to a continuous-time Markov chain. During an active period, packets are generated at regular periods of T_p. We assume that an active period typically consists of multiple consecutive packets ($T_p/T_{on} < 1$), which is the case of practical interest. The on-off source can thus be conveniently viewed as a renewal process with interarrival time distribution given by

$$A(t) = [(1-\frac{T_p}{T_{on}}) + \frac{T_p}{T_{on}}(1-e^{-\frac{(t-T_p)}{T_{off}}})]U(t-T_p), \tag{17}$$

where $U(t)$ is the unit step function, and the Laplace transform

$$A^*(s) = \int_0^\infty e^{-st}dA(t) = [1-\frac{T_p}{T_{on}} + \frac{T_p}{T_{on}(1+sT_{off})}]e^{-sT_p} \tag{18}$$

with the peak packet arrival rate $1/T_p$ and the mean packet arrival rate $1/(T_p+T_pT_{off}/T_{on})$. For CBR traffic, we simply let $T_{on} \to \infty$ and $T_{off} \to 0$ in Equations (17)(18) and obtain

$$A(t) = U(t-T_p) \tag{19}$$

$$A^*(s) = e^{-sT_p}, \tag{20}$$

where T_p now becomes the constant packet interarrival time.

We simulate an 802.11e based wireless LAN with 100 mobile nodes. All nodes are within the transmission range of one another. The channel rate is 2 Mb/s. The traffic parameters are listed as follows.

Voice Traffic (VBR): The voice traffic is modeled as VBR using an *on/off* source with exponentially distributed *on* and *off* periods of 300 ms average each. Traffic is generated during the *on* periods at a rate of 32 kb/s with a packet size of 160 bytes, thus the inter-packet time is 40 ms.

Video Traffic (CBR): The video traffic is modeled as CBR traffic with a rate of 64 kb/s with a packet size of 1000 bytes, thus the inter-packet time is 125 ms.

Similar to [29], we assign the video traffic to AC 2 and the voice traffic to AC 3. Two set of parameters are used to verify the analysis. In setting (a), $AIFS[2] = 60\mu s$, $AIFS[3] = 50\mu s$, $W_{2,0} = 32$, and $W_{3,0} = 16$; in setting (b), $AIFS[2] = 75\mu s$, $AIFS[3] = 50\mu s$, $W_{2,0} = 64$, and $W_{3,0} = 16$. It can be seen that it becomes harder for the video traffic to gain channel

access in setting (b) than in setting (a). In both settings, the number of queues for each traffic class is equal, i.e., $n_2 = n_3$. Note that the network works in the unsaturated case.

Fig. 2(a) and 2(b) respectively illustrate the average delay as a function of the total number of flows, i.e., $n_2 + n_3$ for both settings. In each figure, both the analytical and simulation results are presented. Several observations are made here. First, as the number of flows increases, for either the analytical or simulation results, the delays for both traffic classes increase. The reason is as follows. In the unsaturated case, the collision is not severe with the collision probability less than 0.1, and the queue does not build up. As a result, the queueing delay is small and the MAC layer service time dominates the delay. When the number of competing flows increase, the collision increases and so does the MAC layer service time. Second, the delay for the voice traffic is much smaller than that for the video traffic, which is consistent with the fact that the voice traffic has a higher priority than the video traffic in terms of channel access. Especially, as expected, the delay for the voice traffic in setting (b) is smaller than that in setting (a) and the delay for the video traffic in setting (b) greater than that in setting (a). Third, the G/M/1 and the G/G/1 models deliver very close delay results, both greater than the simulation delays, indicating in practice they can provide the upper bounds for the average delay. As shown later, we use them in the proposed call admission control scheme. We also observe that as the number of flows increases, the gaps between the simulation and analytical results become larger. Nevertheless, our analytical results can serve as upper bounds of the average delay. Finally, it is important to point out that when we keep the network working in the unsaturated case, the delays for both traffic classes are sufficiently small to satisfy their QoS requirements as specified in [15] [16], where the one way transmission delay for interactive communications like VoIP and videoconferencing should be preferably less than 150ms, and must be less than 400ms.

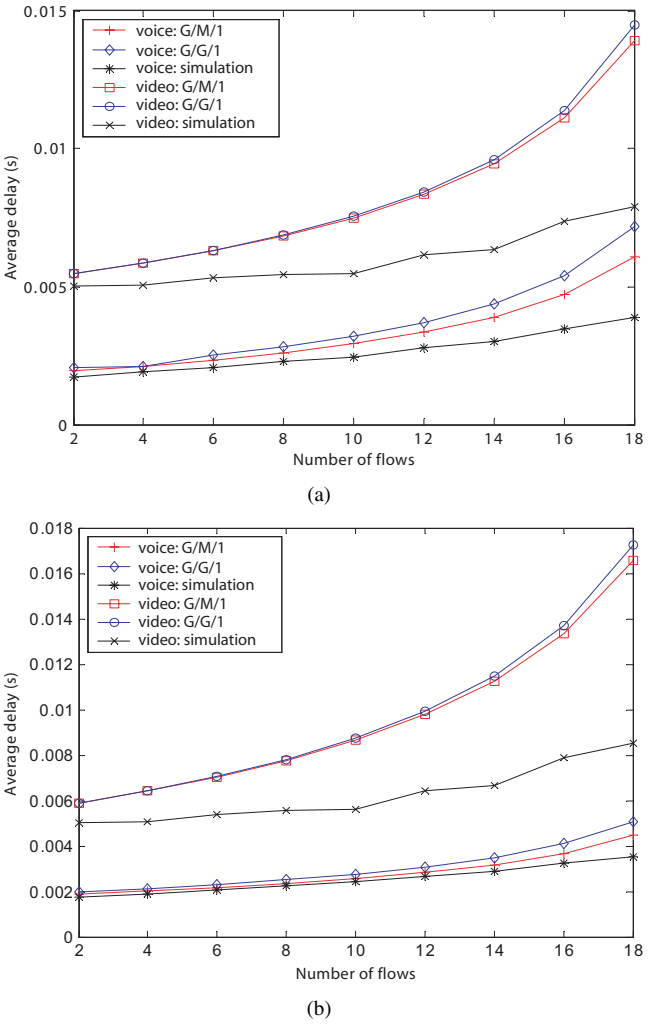


Fig. 2. (a) Average delay (ms) when $AIFS[2] = 60\mu s$, $AIFS[3] = 50\mu s$, $W_{2,0} = 32$, and $W_{3,0} = 16$; (b) average delay (ms) when $AIFS[2] = 75\mu s$, $AIFS[3] = 50\mu s$, $W_{2,0} = 64$, and $W_{3,0} = 16$.

V. Call Admission and Rate Control Algorithm

To keep the network operating in the unsaturated case, where the collision probability is small, the throughput is high, and the delay is short [30], it is crucial to regulate total input traffic. Since the real-time traffic is not greedy in terms of bandwidth usage, and more importantly, has strict delay requirements, call admission control (CAC) is a suitable traffic control mechanism for it. On the other hand, for non-real-time data traffic, which can tolerate delay ranging from seconds to minutes but are greedy in terms of bandwidth usage, rate control (RC) is appropriate. However, for the IEEE 802.11e that relies on contention-based channel access, it is hard to characterize the current traffic conditions. Therefore, we need to find an appropriate control variable for both the admission control and rate control. In the following, we first briefly discuss the concept of channel busyness ratio.

A. Channel Busyness Ratio

The channel busyness ratio, denoted by $r_b \in [0, 1]$, is defined as the portion of the time that the channel is busy in an observation period, which can be directly measured at each node since the IEEE 802.11e MAC is based on carrier sensing. Meanwhile, the channel utilization, denoted by cu, is defined as the portion of the time that the channel is used for successful transmissions in an observation period. Clearly, $cu \leqslant r_b$, and the equality holds only when all the channel busy time is used for successful transmissions. In other words, there is no collision at all.

Because of space limit, we only give several advantages of the channel busyness ratio (refer to [30] for more details). First, as we showed in [30], in the unsaturated case, since the collision probability is very small (typically below 0.1) and the time wasted due to channel collision can be ignored, channel utilization is almost equal to the channel busyness ratio. As a matter of fact, this is also confirmed in Section VI. For this reason, we may use these two terms interchangeably. Moreover, in the unsaturated case, there exists an optimal point where the network achieves the maximum throughput and short delay. The channel busyness ratio at this point changes very little when the number of nodes or the packet lengths change. More details on how to determine the corresponding channel utilization is given in Section V-F. Second, since the EDCA is based on carrier sensing, it is easy to measure the

channel busy time. On the other hand, the channel utilization is not readily measurable as a node cannot distinguish channel collision from channel fading.

Next, we present the CAC and RC schemes in order. For CAC, we present two schemes, namely a comprehensive one and a simplified one.

B. Call Admission Control Scheme I

As specified in the IEEE 802.11e EDCA, the admission control is conducted at the QoS access point (QAP) when the infrastructure mode is used. If the network is working in the ad hoc mode, a mobile node can be elected to coordinate the admission control using one of many algorithms in the literature ([9] [22]). Further discussions on the election algorithm is beyond the scope of this paper. Hereafter, we use the coordinator to denote the QAP or the coordinating node without differentiation. It should be noted that such a coordinator is necessary to avoid the so-called *over-admission* problem, which will occur when several individual nodes, if not coordinated, admit new real-time flows at the same time and cause the admitted traffic to exceed the network capacity.

In the admission control, we should set a quota on the amount of the real-time traffic that the network can admit [7]. Namely, if we measure the traffic amount in terms of its contribution to the channel utilization (or the channel busyness ratio), we should set a quota on the channel utilization that is due to the real-time traffic. We set such a quota, denoted by CU_{rt}, to 80% [1] of the maximum channel utilization, denoted by CU_{max} for two reasons. It first ensures that the best effort traffic is operational all the time, since the best effort traffic is at least entitled to 20% of the channel utilization. In addition, the 20% of the channel utilization for the best effort traffic can be used to accommodate sizable fluctuations caused by the VBR real-time traffic.

In the CAC scheme, three parameters, (R_{mean}, R_{peak}, PK_l), are used to characterize the bandwidth requirement of a real-time flow, where R_{mean} is the average data rate and R_{peak} the peak data rate in (bit/s), and PK_l is the average packet length in bits. For CBR traffic, $R_{mean} = R_{peak}$. For VBR traffic, $R_{mean} < R_{peak}$. When the RTS/CTS mechanism is used, the time associated with a successful transmission, denoted by T_{suc}, is obtained by

$$T_{suc} = RTS + CTS + DATA + ACK + 3SIFS + AIFS, \tag{21}$$

where $DATA$ is the average packet transmission time for the packet of length PK_l. Then, we can calculate the channel utilization cu corresponding to a flow's bandwidth requirement as follows:

$$cu = \mathcal{U}(R) = \frac{R}{PK_l} \times T_{suc}, \tag{22}$$

where $\mathcal{U}$ is the mapping function from the traffic rate to the channel utilization. Thus, a flow's bandwidth requirement can be translated into (cu_{mean}, cu_{peak}), where $cu_{mean} = \mathcal{U}(R_{mean})$ and $cu_{peak} = \mathcal{U}(R_{peak})$.

The coordinator records the total channel utilization due to all admitted real-time flows into two parameters ($cu_{A,mean}$, $cu_{A,peak}$), i.e., the aggregate (cu_{mean}, cu_{peak}). They are updated when a real-time flow joins or leaves. Meanwhile, the coordinator maintains the number of voice flows (AC 3), denoted by n_3, and the number of video flows (AC 2), denoted by n_2.

Before initiating a real-time flow of priority i (i = 2 or 3), a node must send an ADDTS (add traffic stream) request [14] to the coordinator. The ADDTS contains the traffic priority and the traffic specification (TSPEC) corresponding to the specific application, and the TSPEC specifies R_{mean}, R_{peak}, and PK_l (i.e., the nominal MSDU size).

Upon receiving the ADDTS, the coordinator associates the flow with the appropriate AC i and obtains $cu_{i,mean}$ and $cu_{i,peak}$ according to Equation (22). Then, it determines if the flow can be admitted using the following tests:

- First, the remainder of the quota CU_{rt} and CU_{max} should be able to accommodate the new real-time flow, i.e.,
$$\begin{cases} cu_{A,mean} + cu_{i,mean} < CU_{rt} \\ cu_{A,peak} + cu_{i,peak} < CU_{max}. \end{cases} \tag{23}$$
- Second, for each currently existing real-time flow of priority i and the new flow, we can estimate the average delay $\overline{D_i}$ using the G/G/1 model described earlier. It should be less than the delay bound D_i required by the specific application, i.e.,
$$\overline{D_i} \leqslant D_i \quad i = 2, 3. \tag{24}$$

If both of the above conditions are satisfied, the new flow is admitted. The coordinator updates ($cu_{A,mean}$, $cu_{A,peak}$, n_i) accordingly. Otherwise, the new flow is rejected. The coordinator notifies the node of the decision by sending an ADDTS response.

When a real-time flow ends, the source node of the flow should transmit a DELTS (delete traffic stream) containing the TSID (traffic stream identifier) to the coordinator, and the latter updates ($cu_{A,mean}$, $cu_{A,peak}$, n_i) accordingly.

In the above admission control scheme, we should note two points. First, the average delay can be computed offline and stored in a table. Specifically, for each combination of n_i ($i = 2, 3$), the average delay for each traffic class is computed as mentioned earlier. At runtime, the stored values can be looked up without any complex computations. Second, in the above call admission control, we do not consider the effect of the best effort traffic on the delay of the real-time traffic for the following reasons. Since in the 802.11e WLAN, the best effort traffic has a much larger $AIFS$ and contention window CW than the real-time traffic, its effect on the real-time traffic is not as significant as other real-time traffic. More importantly, with the rate control described later, we can further reduce the negative effect.

C. Call Admission Control Scheme II

It can be seen that when making admission decisions, CAC scheme I takes into account both the peak rate and mean rate for the real-time traffic. While this ensures that the network will not be congested in the worst-case scenario, in which all the VBR real-time traffic transmits at its peak rate, it may unnecessarily reject many real-time flows when the ratio

[1]This number is tunable and could be changed depending on the traffic composition in real networks. We choose 80% for our study only.

R_{peak}/R_{mean} is large for some real-time applications. To resolve this problem, we only consider the mean rate in the admission control scheme. Meanwhile, recognizing that it may not be practical for non-QAP nodes to calculate the average delay beforehand if the network works in the ad hoc mode, We further remove the delay test from the admission scheme. Note that this might not be a too bad idea if we consider that as suggested in Section IV, as long as the network is kept working in the unsaturated case and the best effort traffic is well controlled to isolate its effect on the real-time traffic, the delay for the real-time traffic should be small enough to meet the QoS requirements. After making these two changes to CAC scheme I, we get the simplified CAC scheme II as follows.

When a node sends a request with the corresponding TSPEC to the coordinator, the coordinator grants admission if the following test is passed:

- The remainder of the quota CU_{rt} should be able to accommodate the new real-time flow, i.e.,

$$cu_{A,mean} + cu_{i,mean} < CU_{rt}. \tag{25}$$

Note that for CAC scheme II, we implicitly take advantage of the fact that CU_{rt} limits the channel utilization that can be consumed by the real-time traffic, thereby leaving room to accommodate bandwidth fluctuation caused by VBR traffic. But it still requires that the ratio R_{peak}/R_{mean} is not too large.

D. Remarks on Call Admission Control

It can be seen that there exists a tradeoff between strict QoS guarantee and the number of real-time flows that can be accepted, with CAC scheme I targeted for the former and CAC scheme II targeted for the latter. A better balance between these two conflicting objectives could be achieved if better knowledge about the rate-changing pattern of VBR flows is available. In other words, with the aid of such knowledge, we may be able to design another scheme that can accept more real-time flows than CAC scheme I and support better QoS than CAC scheme II. However, it is very hard to obtain precise characterization of real-time traffic a priori given a number of various application scenarios for WLANs. To get around this difficulty, maybe the measurement based admission control approach [17] that was proposed for wired networks can be used. However, it leads to high channel utilization only in the presence of a large number of flows, or a high degree of statistical multiplexing. This might be the case for broadband WLANs with bandwidth 54Mbps or higher. We plan to look into this issue in our future work.

E. Rate Control

The transmission rate of the best effort traffic is controlled based on two criteria. First, the best effort traffic should not affect the QoS level of the admitted real-time traffic. One may argue that this can be easily achieved if the channel access parameters such as $AIFS$ and CW are set much larger than those for the real-time traffic. However, this approach is problematic in that it will unnecessarily impede the best effort traffic from accessing the channel even when there is no heavy real-time traffic in the network, leading to channel underutilization and unreasonably large delay for the best effort traffic. Second, the best effort traffic should be able to promptly access the residual bandwidth left by the real-time traffic in order to efficiently utilize the channel.

Clearly, to meet these criteria, each node needs to accurately estimate the total instantaneous rate of the ongoing real-time traffic. However, this is not an easy task if the network works in the ad hoc mode, where nodes can communicate with one another directly without involving QAP. Meanwhile, even if the network works in the infrastructure mode, since the IEEE 802.11e allows direct links between two non-QAP nodes, all communications may not necessarily go through the QAP. It can thus be concluded that in either mode, there is no node that can accurately monitor all the traffic in the air and control the traffic rate of all the other nodes. Therefore, an effective distributed rate control scheme is desired.

In the rate control scheme, each node needs to monitor the channel busyness ratio r_b during a period of T_{rb}. Let us denote by r_{br} the contribution from the real-time traffic to r_b, and denote by R_{be} the data rate of the best effort traffic at the node under consideration, with the initial value of R_{be} being conservatively set, say one packet per second. The node thus adjusts R_{be} after each T_{rb} according to the following:

$$R_{be_{new}} = R_{be_{old}} \times \frac{CU_{max} - r_{br}}{r_b - r_{br}}, \tag{26}$$

where $R_{be_{new}}$ and $R_{be_{old}}$ are the value of R_{be} after and before the adjustment. Two points are noted on Equation (26). First, we see that the node increases the rate of the best effort traffic if $r_b < CU_{max}$ and decreases the rate otherwise. Second, if all the nodes adjust the rate of its own best effort traffic according to Equation (26), the total best effort data rate will be

$$\sum R_{be_{new}} = \sum R_{be_{old}} \times \frac{CU_{max} - r_{br}}{r_b - r_{br}} \approx \mathcal{U}^{-1}(CU_{max} - r_{br}), \tag{27}$$

where $\sum R_{be_{old}} \approx \mathcal{U}^{-1}(r_b - r_{br})$ is due to the fact that the channel busyness ratio is equal to the channel utilization and $r_b - r_{br}$ is the contribution from the total best effort traffic to r_b. Thus after one control interval T_{rb}, the channel utilization will be approximate to CU_{max}.

To estimation of r_{br}, each mobile node needs to monitor all the traffic in the air. However, to be consistent with the original 802.11e protocol, our scheme only requires mobile nodes to decode the MAC header part, as the original 802.11e does in the NAV procedure. To distinguish real-time packets from best effort packets, we only need to check the most significant bit of the subtype field, which is defined in the IEEE 802.11e as the QoS subfield in data packets. Therefore, the observed channel busyness ratio comprises three pieces of contribution: the contribution from the best effort traffic with a decodable MAC header r_{b1}, that from the real-time traffic with a decodable MAC header r_{b2}, and that of all the traffic with an undecodable MAC header r_{b3} due to collision. So we give an upper bound and a lower bound for r_{br} as follows:

$$r_{b2} \leqslant r_{br} \leqslant r_{b2} + r_{b3}. \tag{28}$$

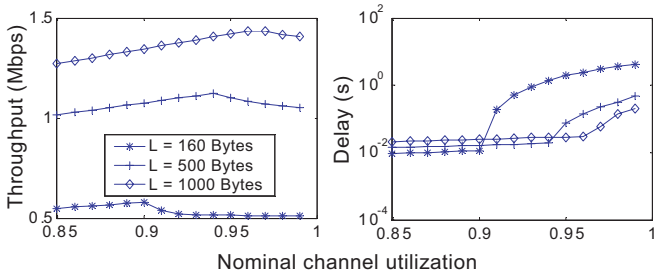


Fig. 3. Choice of CU_{max} for different packet lengths.

To enforce a conservatively increasing and aggressively decreasing law, we thus set r_{br} as follows:

$$r_{br} = \begin{cases} r_{b2}, & if\ r_b < CU_{max} \\ r_{b2} + r_{b3}, & if\ r_b > CU_{max}. \end{cases} \quad (29)$$

We also note that the control interval T_{rb} should be set such that the scheme can be responsive to the change of the channel busyness ratio observed in the air and can smooth out the instantaneous disturbance.

F. Determination of CU_{max}

It is clear that choosing an appropriate maximum channel utilization, CU_{max}, is critical in making both the call admission control and rate control work. Next, we show how to determine an appropriate value for CU_{max}.

First, we consider how the packet length affects the channel utilization. We consider a network of 40 nodes and each node generates CBR traffic. The default 802.11 DCF system parameters are used: $SIFS$ = 10μs, $DIFS$ = 50μs, and the initial $CW = 32$. The RTS/CTS mechanism is used. Fig. 4 shows both the throughput and delay as a function of the input network traffic. Here we use the nominal channel utilization to denote the traffic load, that is, the channel utilization that would result from the traffic load as if there were no collisions at all. We can see that regardless of the packet length, as the traffic load increases, the throughput first increases and then decreases; meanwhile, the delay first increases very slowly and then increases dramatically. In other words, the network enters from unsaturation to saturation. The channel utilization values corresponding to the turning points (or the boundaries between unsaturation and saturation), is CU_{max}. It can also be observed that when the packet length increases, so does CU_{max}. Obviously, to achieve the maximum throughput and short delay, CU_{max} should be set in the range of 0.9 to 0.95.

Second, we consider how robust such a choice of CU_{max} is in the prioritized scenarios. We consider two types of traffic, namely the high priority traffic and low priority traffic. For high priority traffic, $AIFS$ = 50μs and the initial $CW = 16$; the packet length is 500Bytes. For low priority traffic, $AIFS$ = 60μs and the initial $CW = 32$; the packet length is 1000Bytes. Each node generates either a high priority or low priority traffic flow. Fig. 4 shows the throughput and delay when the traffic load increases. Again, we can see that the choice of CU_{max} within the range of [0.9, 0.95] leads to good performance. Note these observations are also true of the case where RTS/CTS is not used.

VI. Performance Evaluation

A. Simulation Configuration

To evaluate the performance, we conduct simulations in OPNET Modeler 10.0 [21]. An 802.11e based wireless LAN with 100 mobile nodes is simulated. All nodes are within the transmission range of one another. In all simulations, channel rate is 2 Mb/s and the RTS/CTS mechanism is used. In addition to the two types of real-time traffic mentioned in section IV, we also consider the greedy best-effort TCP traffic (AC 0), which is of a packet size of 1000 bytes. TCP-Reno is used. So voice, video, and data correspond to AC3, AC2, and AC0 respectively. The $AIFS$ and CW parameters are set as follows. $AIFS[0] = 80\mu s$, $AIFS[2] = 60\mu s$, $AIFS[3] = 50\mu s$; $W_{0,0} = 128$, $W_{2,0} = 32$, and $W_{3,0} = 16$. In such a setting, it is clear that the voice traffic has the highest priority and the TCP traffic has the lowest priority in terms of channel access. $CU_{max} = 0.93$ and $CU_{rt} = CU_{max} * 80\% = 0.744$. The period of measuring the channel busyness ratio $T_{rb} = 2$s. In CAC scheme I, $D_2 = 200ms$ and $D_3 = 100ms$. The simulation time is 120 seconds.

In the simulation, the traffic load is gradually increased. Specifically, a new voice, video or TCP flow is periodically added in an interleaved way, in order to observe how the scheme works and how a newly admitted flow impacts the performance of previously admitted flows. Until 94 seconds, a new voice flow is added at the time instant of $6 \times i$ second $(0 \leqslant i \leqslant 15)$. Likewise, a video flow is added two seconds later and a TCP flow is added 4 seconds later. Furthermore, to simulate the real scenario where the start of real-time flows are randomly spread over time, the start of a voice flow is delayed a random period uniformly distributed in [0ms, 40ms], and that of a video flow delayed a random period uniformly distributed in [0ms, 125ms]. Note that in the simulation period between (94s, 120s], we purposely stop injecting more flows into the network in order to observe how well the scheme performs in a steady state.

B. Simulation Results

1) CAC scheme I and RC: From the simulation results, we find there are a total of 10 voice flows and 10 video flows admitted by 56 seconds; and no more voice or video flows are admitted thereafter. The number of TCP flows increases by one every 6 seconds until 94 seconds. After 94 seconds, as expected, there is no change in the number of flows. This is expected. According to Equation (22), we know that the $cu_{3,mean}$ and $cu_{3,peak}$ for a voice flow are 0.0248 and 0.0496, respectively; and the $cu_{2,mean}$ (=$cu_{2,peak}$) for a video flow is 0.04283. Following the admission criteria in CAC scheme I, after the network admits 10 voice flows and 10 video flows, $cu_{A,mean} = 0.6763$ and $cu_{A,peak} = 0.9243$. Obviously, no more real-time flows can be accepted due to the constraint of $CU_{max} = 0.93$. We should mention that up to 56 seconds, no real-time flows are rejected because the delay criterion specified in Equation (24) cannot be met. During the simulation, neither real-time or best effort packets are lost.

Fig. 5(a) shows the throughput for the three traffic classes throughout the simulation. At the beginning, the TCP traffic

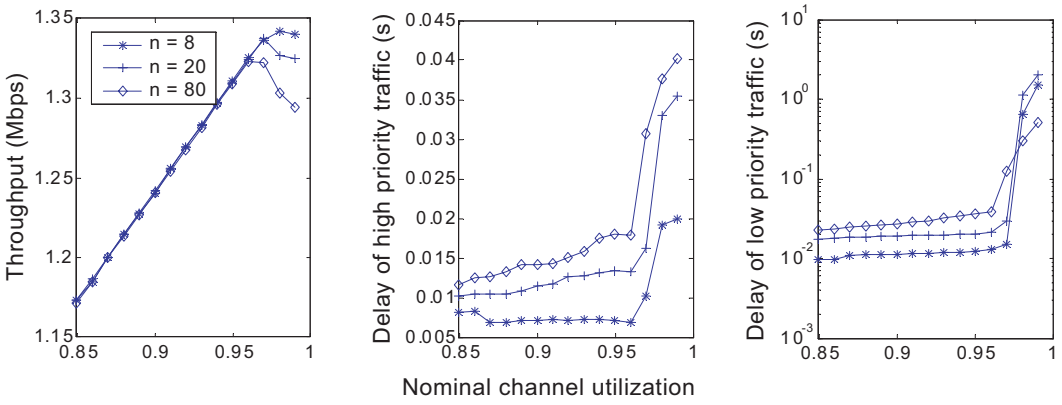


Fig. 4. Choice of CU_{max} for prioritized traffic and different numbers of nodes

has high throughput; then as more real-time flows are admitted, it gradually drops as a result of the rate control as well as the high priority of real-time traffic. Because we set an upper bound CU_{rt} for the real-time traffic, it can be observed that even when the traffic load becomes heavy, TCP traffic, as desired, is not completely starved. Because TCP traffic is allowed to use any available channel capacity left by the real-time traffic, the total channel utilization, namely the sum of the channel utilization due to different types of traffic, stabilizes at as high as 0.9, as shown in Fig. 5(b). Fig. 5(b) also shows that in the unsaturated case, as a result of the very small collision probability, the channel utilization curve coincides with the channel busyness ratio curve.

The packet delay is illustrated in Fig. 5(c), in which every point is averaged over 2 seconds. As expected, it can be observed that the delay for the real-time traffic is kept below 20 ms; moreover, the delay for the voice traffic is much smaller than that for the video traffic. Initially, as the number of admitted real-time flows increases, the delay increases. Note that the increase of delay is not due to the TCP traffic, but mainly due to the increasing number of competing real-time flows. Then, the delay oscillates around a stable value. Fig. 5(d) presents the delay distribution for the voice and video traffic without any averaging. More detailed statistics of delay and delay variation are given in Table I. Again, no averaging is taken. As shown in Table I, the 97 percentile delay values for voice and video are 18.5 ms and 29.2 ms respectively, and the 99 percentile delay values for voice and video are 24.6 ms and 37.1 ms respectively. It is known that for the real-time traffic, packets that fail to arrive in time are simply discarded. Given the allowable $1\% \sim 3\%$ packet loss rate, these delays are well within the bounds given in [15] [16]. The good delay performance indicates that CAC scheme I and RC together can effectively guarantee the delay and delay jitter requirements of the real-time traffic, even in the presence of highly dynamic TCP traffic.

2) CAC scheme II and RC: Unlike the previous case, when CAC scheme II and RC are used, we observe that there are a total of 11 voice flows and 11 video flows admitted by 62 seconds; and no more voice or video flows are admitted thereafter. Again, the number of TCP flows increases by one every 6 seconds until 94 seconds. After 94 seconds, there is

TABLE I
THE MEAN, STANDARD DEVIATION (SD), AND 97'TH, 99'TH, 99.9'TH PERCENTILE DELAYS (S) FOR VOICE AND VIDEO WHEN CAC SCHEME I AND RC ARE USED.

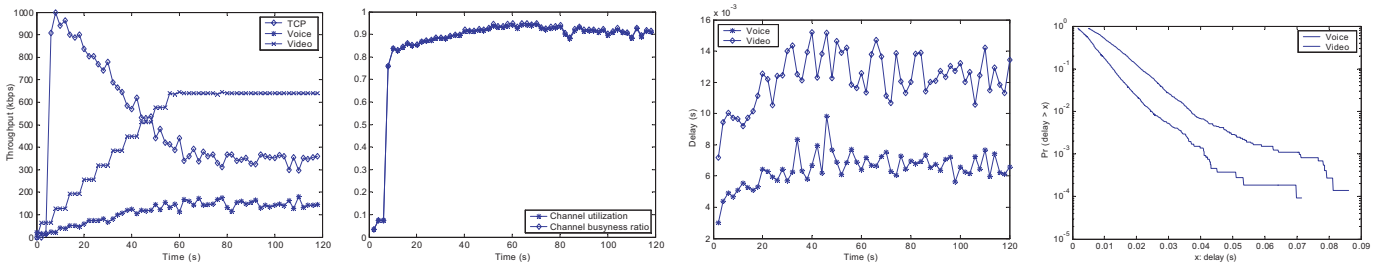
	mean	SD	97 %ile	99 %ile	99.9 %ile
VBR Voice	0.0065	0.0051	0.0185	0.0246	0.0411
CBR Video	0.0123	0.0074	0.0292	0.0371	0.0708

no change in the number of flows. The reason that more real-time flows are admitted in this case is the following. In the previous case, after 10 voice flows and 10 video flows are admitted, $cu_{A,peak}$ is close to CU_{max} and thus no more real-time flows can be accepted. Since CAC scheme II eliminates that constraint, now only the constraint CU_{rt} works. After 11 voice flows and 11 video flows get into the network, $cu_{A,mean}$ is equal to 0.7439 and close to CU_{rt}. Thus, no more real-time flows can be admitted. During the simulation, neither real-time or best effort packets are lost.

In Fig. 6(a), we see as one more voice flow and one more video flow are accepted compared to the previous case, the TCP throughput in the steady state drops by a corresponding amount. The channel utilization also remains steadily high except that some slight fluctuations are observed as opposed to that in previous case, since more VBR voice flows in the network. Fig. 6(c) and 6(d) demonstrate that the delay requirements of the real-time traffic can be adequately met. However, as expected, the results are a bit worse than those in the previous case. This can also be seen in Table II, where the 97 percentile, 99 percentile, and 99.9 percentile delay values for voice and video slightly increase. As a whole, however, the good performance in terms of both throughput and delay indicates that this simplified CAC scheme II in combination with RC still works well.

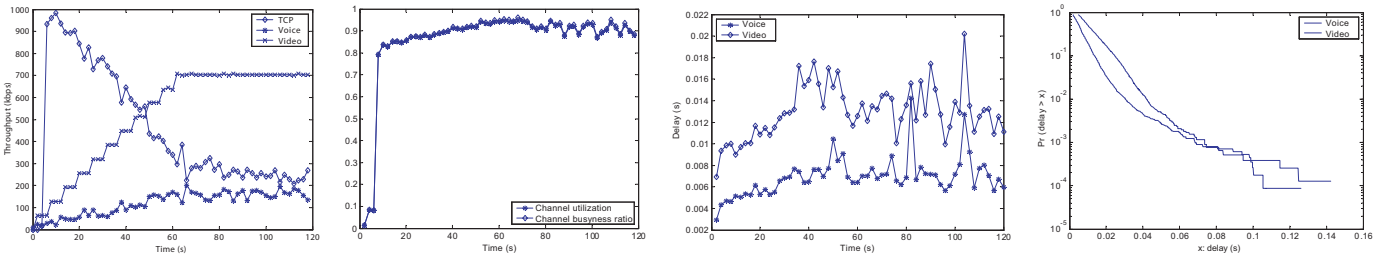
VII. CONCLUSION

While the emerging IEEE 802.11e wireless LAN supports prioritized services, it cannot provide strict QoS for the real-time traffic. In this paper, we enhance the 802.11e by



(a) (b) (c) (d)

Fig. 5. (a) Aggregate throughput, (b) channel busyness ratio and channel utilization, (c) average delay of voice and video traffic, (d) delay distribution of voice and video traffic.



(a) (b) (c) (d)

Fig. 6. (a) Aggregate throughput, (b) channel busyness ratio and channel utilization, (c) average delay of voice and video traffic, (d) delay distribution of voice and video traffic.

TABLE II
THE MEAN, STANDARD DEVIATION (SD), AND 97'TH, 99'TH, 99.9'TH PERCENTILE DELAYS (S) FOR VOICE AND VIDEO WHEN CAC SCHEME II AND RC ARE USED.

	mean	SD	97 %ile	99 %ile	99.9 %ile
VBR Voice	0.0069	0.0066	0.0209	0.0306	0.0684
CBR Video	0.0130	0.0089	0.0338	0.0421	0.0738

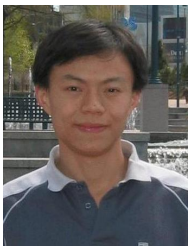
proposing two call admission schemes and a rate control scheme. We first build an analytical model to analyze the average delay for the traffic with different priorities and derive an estimate, which is then used in the call admission control mechanism. The analytical results show the 802.11e WLAN can satisfy the delay requirements of the real-time traffic as long as the network is tuned to operate in the unsaturated case. Then, relying on the novel use channel busyness ratio, we demonstrate that the two call admission control schemes ensure QoS guarantees for the real-time traffic and the rate control scheme allows the best effort traffic to use the residual channel capacity left by the real-time traffic. Finally, the simulation results show that the proposed schemes successfully guarantee stringent QoS requirements of real-time services, while achieving high channel utilization.

REFERENCES

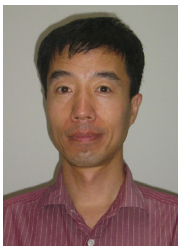
[1] I. Ada and C. Castelluccia, "Differentiation mechanisms for IEEE 802.11," in *Proc. IEEE INFOCOM'01*, Anchorage, Alaska, April 2001.
[2] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a media access protocol for wireless LAN's," in *Proc. ACM SIGCOMM 1994*.
[3] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.
[4] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 protocol: design and performance evaluation of an adaptive backoff mechanism," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 9, pp. 1774–1786, Sept. 2000.
[5] F. Cali, M. Conti, and E. Gregori, "Tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. Networking*, vol. 8, no. 6, pp. 785–799, Dec. 2000.
[6] S. Choi, J. Prado, S. Mangold, and S. Shankar, "IEEE 802.11e contention-based channel access (EDCF) performance evaluation," in *Proc. IEEE ICC'03*, May 2003.
[7] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time application in an integrated services packet network: architecture and mechanism," in *Proc. of ACM SIGCOMM*, 1992.
[8] C. H. Foh and M. Zukerman, "Performance analysis of the IEEE 802.11 MAC protocol," in *Proc. European Wireless 2002*, Florence, Italy, Feb. 2002.
[9] H. Garcia-Molina, "Elections in a distributed computing system," *IEEE Trans. Comput.*, vol. 31, no. 1, Jan. 1982.
[10] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory, 3rd ed.* John Wiley & Sons, Inc, 1998.
[11] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 856–868, Sept. 1986.
[12] T. S. Ho and K. C. Chen, "Performance analysis of IEEE 802.11 CSMA/CA medium access control protocol," in *Proc. IEEE PIMRC 1996*.
[13] *IEEE standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, ISO/IEC 8802-11: 1999(E), 1999.
[14] *Draft Supplement to Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)*, IEEE Std 802.11e/D8.0, Feb. 2004.
[15] ITU-T G.114. One-way transmission time, 1996.
[16] ITU-T G.1010. End-user multimedia QoS categories, 2001.
[17] S. Jamin, P. B. Danzig, S. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated service packet networks," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, pp. 56–70, Feb. 1997.
[18] L. Kleinrock, *Queueing Systems, volume I*. John Wiley & Sons, 1975
[19] L. Kleinrock, *Queueing Systems, volume II*. John Wiley & Sons, 1975

[20] S. Mangold, S. Choi, P. May, O. Klein, G. Hietz, and L. Stibor, "IEEE 802.11e wireless LAN for quality of service," in *Proc. European Wireless'02*, Florence, Italy, Feb. 2002.
[21] OPNET Modeler 10.0. http://www.opnet.com.
[22] S. Singh and J. Kurose, "Electing 'good' leaders," *J. Par. Distr. Comput.*, vol. 18, no. 1, May 1993.
[23] J. L. Sobrinho and A. S. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," *Bell Labs Tech. J.*, Autumn 1996.
[24] A. Veres, A. T. Campbell, M. Barry, and L.-H. Sun, "Supporting service differentiation in wireless packet networks using distributed control," *IEEE J. Sel. Area Commun.*, vol. 19, no. 10, pp. 2081–2093, Oct. 2001.
[25] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *Proc. IEEE INFOCOM'02*, New York, June 2002.
[26] Y. Xiao, "Enhanced DCF of IEEE 802.11e to support QoS," in *Proc. IEEE WCNC'03*, New Orleans, Louisiana, March 2003.
[27] Y. Xiao, H. Li, and S. Choi, "Protection and guarantee for voice and video traffic in IEEE 802.11e Wireless LANs," in *Proc. IEEE INFOCOM'04*, Hong Kong, China, March 2004.
[28] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *J. Wireless Communications and Mobile Computing*, vol. 4, pp. 917-931, Dec. 2004.
[29] X. Chen, H. Zhai, and Y. Fang, "Enhancing the IEEE 802.11e in QoS support: analysis and mechanisms," in *Proc. Second International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine'05)*, Aug. 2005.
[30] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service," *IEEE Trans. Wireless Commun.*, vol. 4, no.6, pp. 3084-3094, Nov. 2005.

Xiang Chen (S'03-M'05) received his Ph.D. degree in electrical and computer engineering from the University of Florida in 2005, and received his M.E. and B.E. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2000 and 1997, respectively. He is now a senior research engineer with Motorola Labs. His research interests include resource management, medium access control, and QoS in wireless networks. He is a member of Tau Beta Pi.

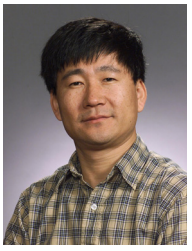


Hongqiang Zhai (S'03) received the B.E. and M.E. degrees in electrical engineering from Tsinghua University, Beijing, China, in July 1999 and January 2002 respectively. He worked as a research intern in Bell Labs Research China from June 2001 to December 2001, and in Microsoft Research Asia from January 2002 to July 2002. Currently he is pursuing the PhD degree in the Department of Electrical and Computer Engineering, University of Florida. He is a student member of ACM and IEEE.



Xuejun Tian graduated from Hebei University, China in 1985. He received his MS degree from Department of Electrical and Mechanical Engineering, Tianjin Institute of Technology, China in 1991, and Ph.D. degree from Department of Intelligence and Computer Science, Nagoya Institute of Technology, Japan, in 1998, respectively. Since 1998, he has been an assistant professor in Department of Information Systems, Faculty of Information Science and Technology, Aichi Prefectural University, Japan. From July 2003 to June 2004, he was a Visiting Assistant Professor in Department of Electrical and Computer Engineering at University of Florida, Gainesville, FL.

Dr. Tian is a member of the IEICE (the Institute of Electronics, Information and Communication Engineers, Japan) and a member of the IEEJ (the Institute of Electrical Engineers of Japan). His research interests include QoS, wireless networks, mobile communications and ubiquitous computing.



Yuguang Fang (S'92-M'94-S'96-M'97-SM'99) received a Ph.D degree in Systems and Control Engineering from Case Western Reserve University in January 1994, and a Ph.D degree in Electrical Engineering from Boston University in May 1997. From July 1998 to May 2000, he was an Assistant Professor in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology. In May 2000, he joined the Department of Electrical and Computer Engineering at University of Florida where he got the early promotion with tenure in August 2003 and became a full Professor in 2005. He has published over 160 papers in refereed professional journals and conferences. He received the National Science Foundation Faculty Early Career Award in 2001 and the Office of Naval Research Young Investigator Award in 2002. He is currently serving as an Editor for many journals, including *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Mobile Computing*, and *ACM Wireless Networks*.