Enhancing the IEEE 802.11e in QoS Support: Analysis and Mechanisms*

Xiang Chen, Hongqiang Zhai, and Yuguang Fang
Department of Electrical & Computer Engineering
University of Florida, Gainesville, Florida 32611
{xchen@ecel, zhai@ecel, fang@ece}.ufl.edu

Abstract

Despite its support of prioritized services, the IEEE 802.11e Enhanced Distributed Channel Access (EDCA) 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 upper bounds for both delay means and variations for services of different priorities in the non-saturated 802.11e WLAN, showing that the QoS requirements of real-time services can be satisfied if the input traffic is properly regulated. Based on the analysis, we then propose a call admission control scheme and a rate control scheme to ensure that QoS requirements of real-time services are statistically guaranteed and that best effort services can efficiently use the residual bandwidth.

1 INTRODUCTION

In order to support quality of service (QoS) in the IEEE 802.11 Distributed Coordination Function (DCF) [8] that only provides best effort services in its current form, 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 [9] [4]. In EDCA, traffic of different priorities is assigned to one of four transmit queues, which respectively correspond to four access categories (ACs). Each AC 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 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 $AIFS[i] \geqslant AIFS[j]$. 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 AC, which performs like a single node in the DCF.

The creation of the EDCA is due to extensive research works that aimed to support prioritized service over the 802.11 DCF [1] [14]. Despite providing prioritized QoS, the EDCA still cannot support strict QoS for real-time applications like voice and video as shown in [4] [16]. Recently, we proposed a call admission and rate control scheme for the 802.11 DCF to provide statistical QoS guarantee [18]; however, it can only support uniform QoS for all services.

Meanwhile, considerable effort was devoted to the theoretical analysis of the 802.11 DCF [2] [3] [6] [15] [17]. Recently, in [12], Kong et al. studied the performance of the 802.11e in the saturated case. Nevertheless, no analysis were focused on the performance of the EDCA in the non-saturated case.

We have found in [7] that it is in the non-saturated case that the 802.11 achieves the maximum throughput and small delay because of the low collision probability. Motivated by this discovery, we aim to tune the network to work in the non-saturated case. In this paper, we make the following contributions. First, we build an analytical model to derive the upper bounds of delay means and variations for the traffic of different priorities in the non-saturated 802.11e 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. Second, we propose a call admission and rate control framework based on the novel use of the channel busyness ratio, which is easy to obtain and can accurately represent the network status. By utilizing the derived mean delay upper bound, the call admission control ensures the QoS requirements of the real-time traffic are met. 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.

*This work was supported in part by the U.S. National Science Foundation under Faculty Early Career Development Award ANIR0093241 and by the U.S. Office of Naval Research under Young Investigator Award N000140210464.

Proceedings of the 2nd Int'l Conf. on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine'05)
0-7695-2423-0/05 $20.00 © 2005 IEEE

The remainder of this paper is organized as follows. The upper bounds are derived and verified in Section 2. We then present our proposed call admission and rate control scheme in Section 3. In Section 4, the performance is evaluated through comprehensive simulation studies. Finally, Section 5 concludes this paper.

2 DERIVATION OF THE UPPER BOUNDS

This section focuses on the delay analysis of the IEEE 802.11e EDCA in the non-saturated case. We consider the case where the RTS/CTS mechanism is used and our analysis can also be applied 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 ACs in each active nodes. Let i (= 0, 1, 2, 3) denote the priority of the four ACs, with $i = 3$ being the highest priority. Also, let n_i denote the number of ACs of priority i in the network. Each AC is treated as an independent node.

2.1 Markov Chain Model for the IEEE 802.11e

Consider a priority i AC. We define $b(i,t)$ as a stochastic process representing the value of the backoff counter k at time t, and $s(i,t)$ as a stochastic process representing the backoff stage j at time t, 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

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

Let p_i denote the probability of collision seen by a transmitted packet from an AC i. Similar to [2] [7], 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.

By solving this Markov chain, we obtain τ_i, the probability that a node of priority i transmits in a random slot given that the queue is not empty as shown in Equation (2). Once τ_i is known, p_i can be obtained as shown in Equation (3), where $P_{i,0}$ is the probability that the transmit queue of an AC i is empty.

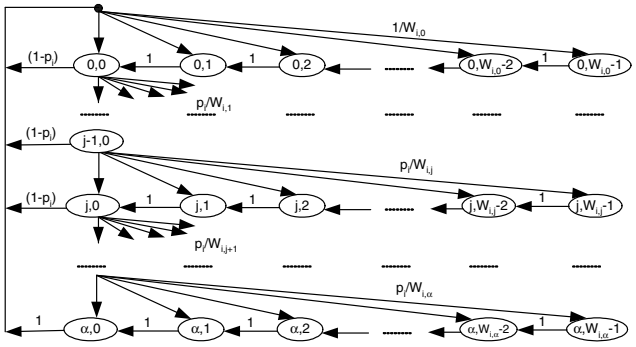


Figure 1. Markov chain for the 802.11e backoff

2.2 Probability Distribution of MAC Service Time

Next, we derive the probability distribution of the MAC service time for a priority i AC using the method proposed in [17]. The *MAC service time* for a priority i packet, denoted by Ts_i, 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 either successfully transmitted or dropped after α times of failed transmissions. Since there exists a one-to-one correspondence between the probability generating function (PGF), i.e., the Z-transform of the probability distribution function denoted by $G_i(Z)$, and the probability distribution of the MAC service time, we choose to calculate the PGF first.

As described in Section 2.1, given the collision probability p_i, we can model the backoff process with a Markov chain. In this chain model, if the PGF of the state transfer time between two states is known, then we can obtain the PGF of the MAC service time. Specifically, if we denote the PGFs of a collision period, a successful transmission period, and the decrement of the backoff timer as $C_i(Z)$, $S_i(Z)$, and $D_i(Z)$, respectively, we can transform the chain model in Fig. 1 into the PGF diagram shown in Fig. 2.

Now we describe how to calculate $C_i(Z)$, $S_i(Z)$, and $D_i(Z)$. Since the collision period associated with a priority i AC is $RTS + SIFS + CTS + AIFS[i]$, we can obtain $C_i(Z)$ as

$$C_i(Z) = Z^{RTS+SIFS+CTS+AIFS[i]} \quad (4)$$

Similarly, we can obtain $S_i(Z)$ as

$$S_i(Z) = Z^{RTS+CTS+3SIFS+DATA_i+ACK+AIFS[i]} \quad (5)$$

where $DATA_i$ is the average packet transmission time in a successful transmission period for AC i.

To calculate $D_i(Z)$, we need to examine how the backoff timer varies. After an idle time slot, denoted by σ, it

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

$$p_i = 1 - \left[\prod_{l=0}^{i-1}(1-(1-P_{l,0})\tau_l)^{n_l}\right](1-(1-P_{i,0})\tau_i)^{n_i-1}\left[\prod_{l=i+1}^{3}(1-(1-P_{l,0})\tau_l)^{n_l}\right] \quad (3)$$

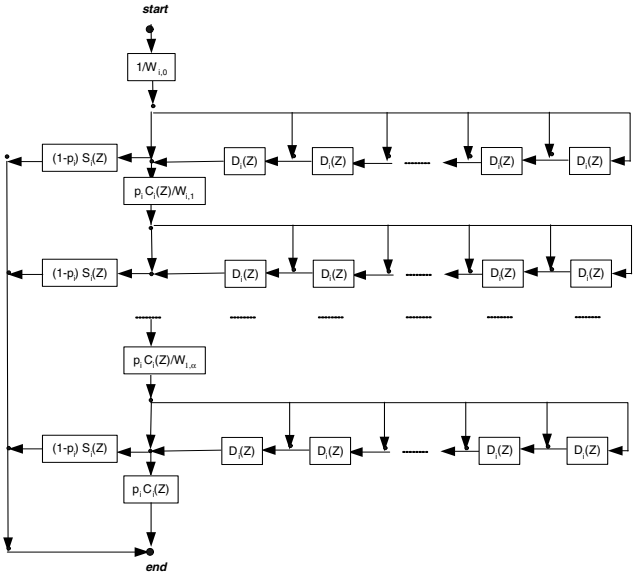


Figure 2. PGF diagram for the backoff

will decrease by 1; after a collision period or successful transmission period, it will stay unchanged. This successful transmission period can be characterized by $S_i'(Z)$:

$$S_i'(Z) = Z^{RTS+CTS+3SIFS+\overline{DATA}+ACK+AIFS[i]} \quad (8)$$

where $\overline{DATA}$ is the average packet transmission time in a successful transmission period for all the ACs except the AC i under consideration. Denote by p_s the probability that among those ACs, there is one AC that transmits successfully. Clearly, it can be obtained as shown in Equation (6). Then, $\overline{DATA}$ can be obtained as shown in Equation (7).

Then, $D_i(Z)$ can be obtained as

$$D_i(Z) = \frac{(1-p_i)Z^{\sigma}}{1-p_s S_i'(Z)-(p_i-p_s)C_i(Z)} \quad (9)$$

We then can use the Mason formula to solve for the transfer function from the "start" point to the "end" point, i.e., the PGF of the MAC service time $G_i(Z)$ as follows.

$$G_i(Z) = (1-p_i)S_i(Z)\sum_{j=0}^{\alpha}\left[[p_iC_i(Z)]^j\prod_{k=0}^{j}H_k(Z)\right] + [p_iC_i(Z)]^{\alpha+1}\prod_{k=0}^{\alpha}H_k(Z) \quad (10)$$

where

$$H_k(Z) = \begin{cases} \sum_{l=0}^{2^kW_{i,0}-1}\frac{D_i^l(Z)}{2^kW_{i,0}}, & (0 \leqslant k \leqslant m) \\ \sum_{l=0}^{2^mW_{i,0}-1}\frac{D_i^l(Z)}{2^mW_{i,0}}, & (m \leqslant k \leqslant \alpha) \end{cases} \quad (11)$$

Note that $G_i(Z)$ is a function of the collision probability p_i. Once we obtain $G_i(Z)$, both the mean and variation of the MAC service time can be derived by taking the derivative with respect to Z:

$$\begin{cases} \frac{1}{\mu_i} = G_i'(Z)|_{Z=1} \\ \sigma_i^2 = G_i''(Z)|_{Z=1} + G_i'(Z)|_{Z=1} - (G_i'(Z)|_{Z=1})^2 \end{cases} \quad (12)$$

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

$$P_{i,0} = 1 - \frac{\lambda_i}{\mu_i} \quad (13)$$

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 Equation (2)(3)(13) 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, $G_i(Z)$ is also completely determined.

2.3 Upper Bound of the Average Delay and Delay Variation

The delay that a packet belonging to AC i experiences, denoted by T_i, can be expressed as follows:

$$T_i \approx Ts_i + R_i \quad (14)$$

where R_i is the residual MAC service time seen by the packet under consideration. Note that in the above equation,

$$p_s = \sum_{j=0,j\neq i}^{3}\left[n_j(1-P_{j,0})\tau_j(1-(1-P_{j,0})\tau_j)^{n_j-1}(1-(1-P_{i,0})\tau_i)^{n_i-1}\prod_{k=0,k\neq i,j}^{3}(1-(1-P_{k,0})\tau_k)^{n_k}\right] + (n_i-1)(1-P_{i,0})\tau_i(1-(1-P_{i,0})\tau_i)^{n_i-2}\prod_{k=0,k\neq i}^{3}(1-(1-P_{k,0})\tau_k)^{n_k} \quad (6)$$

$$\overline{DATA} = \frac{1}{p_s}\sum_{j=0,j\neq i}^{3}\left[DATA_j n_j(1-P_{j,0})\tau_j(1-(1-P_{j,0})\tau_j)^{n_j-1}(1-(1-P_{i,0})\tau_i)^{n_i-1}\prod_{k=0,k\neq i,j}^{3}(1-(1-P_{k,0})\tau_k)^{n_k}\right] + \frac{1}{p_s}DATA_i(n_i-1)(1-P_{i,0})\tau_i(1-(1-P_{i,0})\tau_i)^{n_i-2}\prod_{k=0,k\neq i}^{3}(1-(1-P_{k,0})\tau_k)^{n_k} \quad (7)$$

when the packet arrives, the queue is assumed to be empty except the packet currently being served. This is the case in the non-saturated case [7]. Since the probability distribution of the MAC service time is known, using the Residual Life Theorem [11], we can easily obtain both the mean and variation of R_i:

$$\begin{aligned} E[R_i] &= \frac{P_b E[Ts_i^2]}{2E[Ts_i]} \\ VAR[R_i] &= \frac{P_b E[Ts_i^3]}{3E[Ts_i]} - (\frac{P_b E[Ts_i^2]}{2E[Ts_i]})^2 \end{aligned} \quad (15)$$

where P_b is the probability that the server is busy when the packet arrives.

Then, the upper bounds of the mean and variation of the MAC service time can be obtained by using the fact that P_b belongs to $[0, 1]$,

$$\begin{aligned} E[T_i] &\approx E[Ts_i] + E[R_i] \leqslant E[Ts_i] + \frac{E[Ts_i^2]}{2E[Ts_i]} \\ VAR[T_i] &\approx VAR[Ts_i] + VAR[R_i] \\ &\leqslant VAR[Ts_i] + \frac{5E[Ts_i^3]}{12E[Ts_i^3]} - (\frac{E[Ts_i^2]}{2E[Ts_i]})^2 \end{aligned} \quad (16)$$

2.4 Model Validation

In this section, we validate our analytical results through simulations. We simulate an 802.11e based wireless LAN. All nodes are within the transmission range of one another. The channel rate is 2 Mb/s.We consider two kinds of real-time traffic, i.e., VBR voice traffic and CBR video traffic. The traffic parameters are listed as follows.

VBR Voice Traffic: 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.

CBR Video Traffic: a constant rate of 64 kb/s with a packet size of 1000 bytes. The inter-packet time is 125 ms.

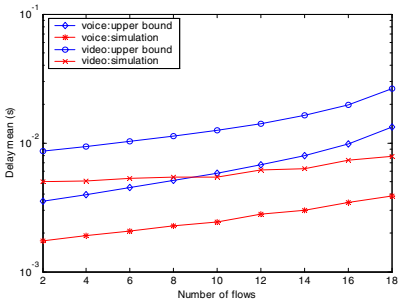
According to the 802.11e [9], we assign the video traffic to AC 2 and the voice traffic to AC 3. $AIFS[2] = 60\mu s$, $AIFS[3] = 50\mu s$, $W_{2,0} = 32$, and $W_{3,0} = 16$. The number of flows for each traffic class is equal, i.e., $n_2 = n_3$. Note that the network works in the non-saturated case.

Fig. 3(a) and 3(b) respectively illustrate the delay mean and standard deviation as a function of the total number of flows, i.e., $n_2 + n_3$. In each figure, both the analytical and simulation results are presented. Several observations are made. First, as the number of flows increases, for either the analytical or simulation results, the delays and standard deviations for both traffic classes increase as a result of the increasing collision level. 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 short packet size and a higher priority than the video traffic in terms of channel access. Third, the upper bounds for both the mean and variation hold, indicating that they can be used in the proposed call admission control scheme presented below.

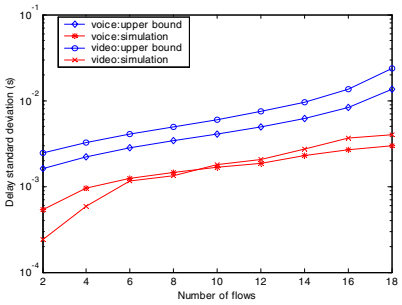
Finally, it is important to point out that when we keep the network working in the non-saturated case, the delays for both traffic classes are sufficiently small to satisfy their QoS requirements as specified in [19] [20], where the one way transmission delay for interactive communications like VoIP and videoconferencing should be preferably less than 150ms, and must be less than 400ms.

3 CALL ADMISSION AND RATE CONTROL ALGORITHM

To keep the network operating in the non-saturated case, we propose call admission control (CAC) for real-time traffic and rate control (RC) for best-effort traffic. To characterize the current traffic conditions, we use the concept of channel busyness ratio ([7]). 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. When the collision level in the network is low, as is in the non-saturated case, the channel busyness ratio is almost the same as the channel utilization, which, denoted by u, is defined as the portion of the time that the channel is used for successful transmissions in an observation period.



(a)



(b)

Figure 3. (a) Delay mean, (b) delay standard deviation; $AIFS[2] = 60\mu s$**,** $AIFS[3] = 50\mu s$**,** $W_{2,0} = 32$ **and** $W_{3,0} = 16$**.**

3.1 Call Admission Control

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. Hereafter, we use the coordinator to denote the QAP or the coordinating node without differentiation.

We should set a quota on the channel utilization that is due to the real-time traffic [5]. We set such a quota, denoted by U_{rt}, to 80% [1] of the maximum channel utilization, denoted by U_{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 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:

$$T_{suc} = RTS + CTS + DATA + ACK + 3SIFS + AIFS \tag{17}$$

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

$$u = \mathcal{U}(R) = \frac{R}{L} \times T_{suc} \tag{18}$$

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 (u_{mean}, u_{peak}), where $u_{mean} = \mathcal{U}(R_{mean})$ and $u_{peak} = \mathcal{U}(R_{peak})$.

The coordinator records the total channel utilization due to all admitted real-time flows into two parameters ($u_{A,mean}$, $u_{A,peak}$), i.e., the aggregate (u_{mean}, u_{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 [9] 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 $u_{i,mean}$ and $u_{i,peak}$ according to Equation (18). Then, it determines if the flow can be admitted using the following tests:

- First, the remainder of the quota U_{rt} and U_{max} should be able to accommodate the new real-time flow, i.e.,

$$\begin{cases} u_{A,mean} + u_{i,mean} < U_{rt} \\ u_{A,peak} + u_{i,peak} < U_{max} \end{cases} \tag{19}$$

- Second, for each currently existing real-time flow of priority i and the new flow, the analytically derived delay upper bound, denoted by $D_{ub,i}$, should be no greater than the delay bound D_i required by the specific application, i.e.,

$$D_{ub,i} \leqslant D_i \quad i = 2, 3 \tag{20}$$

If both of the above conditions are satisfied, the new flow is admitted. The coordinator updates ($u_{A,mean}$, $u_{A,peak}$,

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

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 ($u_{A,mean}$, $u_{A,peak}$, n_i) accordingly.

In the above admission control scheme, we should note two points. First, the analytical delay upper bound can be computed offline and stored in a table for each combination of n_i ($i = 2, 3$). 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.

3.2 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:

$$R_{be_{new}} = R_{be_{old}} \times \frac{U_{max} - r_{br}}{r_b - r_{br}} \tag{21}$$

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 (21). First, we see that the node increases the rate of the best effort traffic if $r_b < U_{max}$ and decreases the rate otherwise. Second, if all the nodes adjust the rate of its own best effort traffic according to Equation (21), the total best effort data rate will be

$$\sum R_{be_{new}} = \sum R_{be_{old}} \times \frac{U_{max} - r_{br}}{r_b - r_{br}} \approx \mathcal{U}^{-1}(U_{max} - r_{br}) \tag{22}$$

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 U_{max}.

To estimate r_{br}, each mobile node needs to monitor all the traffic in the air by decoding 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. 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.

4 PERFORMANCE EVALUATION

4.1 Simulation Configuration

To evaluate the performance, we conduct simulations in OPNET Modeler 10.0 [13]. An 802.11e based wireless LAN 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 2.4, we also consider the greedy best-effort TCP traffic (AC 0), which is of a packet size of 1000 bytes. 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. $U_{max} = 0.93$ and $U_{rt} = U_{max} * 80\% = 0.744$. The period of measuring the channel busyness ratio $T_{rb} = 2$s. $D_2 = 200ms$ and $D_3 = 100ms$. The simulation time is 120 seconds.

In the simulation, 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. Note that in the simulation period between (94s, 120s], we purposefully stop injecting more flows into the network in order to observe how well the scheme performs in a steady state.

4.2 Simulation Results

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 (18), we know that the $u_{3,mean}$ and $u_{3,peak}$ for a voice flow are 0.0248 and 0.0496, respectively; and the $u_{2,mean}$ (=$u_{2,peak}$) for a video flow is 0.04283. Following the admission criteria in the CAC scheme, after the network admits 10 voice flows and 10 video flows, $u_{A,mean} = 0.6763$ and $u_{A,peak} = 0.9243$. Obviously, no more real-time flows can be accepted due to the constraint of $U_{max} = 0.93$. We should mention that up to 56 seconds, no real-time flows are rejected because the delay criterion specified in Equation (20) cannot be met. During the simulation, neither real-time or best effort packets are lost.

Fig. 4(a) shows the throughput for the three traffic classes throughout the simulation. At the beginning, the TCP traffic has high throughput; then as more real-time flows are admitted, it gradually drops as a result of the rate control. Because we set an upper bound U_{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. 4(b). Fig. 4(b) also shows that in the non-saturated case, as a result of the very small collision probability, the channel utilization curve coincides with the channel busyness ratio curve.

The end-to-end delay is illustrated in Fig. 4(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. More detailed statistics of delay and delay variation are given in Table 1, where no averaging is taken. The good delay performance indicates that CAC 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.

Table 1. The mean, standard deviation (SD), and 97'th, 99'th, 99.9'th percentile delays (s) for voice and video.

	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

5 CONCLUSION

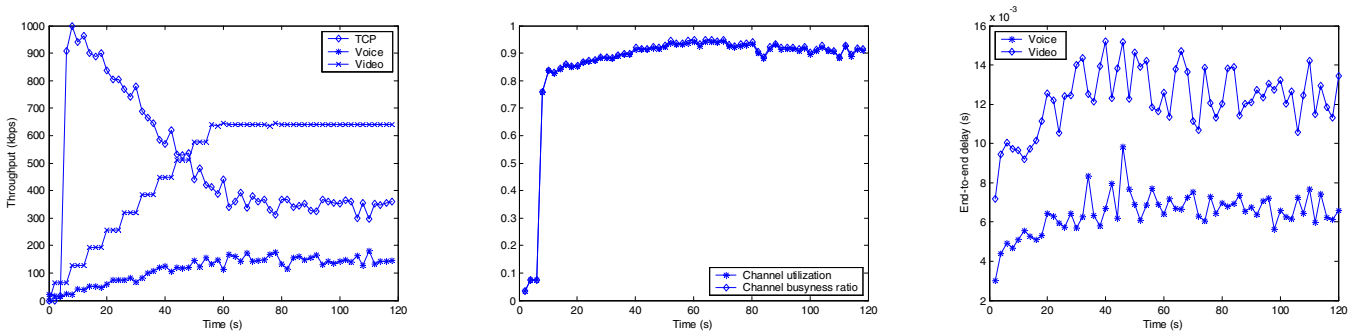
In this paper, we enhance the 802.11e in supporting QoS by proposing a call admission scheme and a rate control scheme. We first analytically derived the upper bounds of delay means and variations for the traffic with different priorities, which are then used in the call admission control mechanism. 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 non-saturated case. Then, using the channel busyness ratio, we demonstrated that the call admission control scheme guarantees QoS for the real-time traffic and the rate control scheme allows the best effort traffic to use the residual channel capacity efficiently. The simulation results verified the performance of the proposed schemes.

References

[1] I. Ada and C. Castelluccia, "Differentiation mechanisms for IEEE 802.11," in *IEEE INFOCOM'01*, Anchorage, Alaska, April 2001.

[2] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE JSAC*, vol. 18, no. 3, March 2000.

[3] F. Cali, M. Conti, and E. Gregori, "Tuning of the IEEE 802.11 protocol to achieve a theoretical throughput



(a) (b) (c)

Figure 4. (a) Aggregate throughput, (b) channel busyness ratio and channel utilization, (c)average end-to-end delay of voice and video traffic.

limit," *IEEE/ACM Transactions on Networking*, vol. 8, no. 6, Dec. 2000.

[4] S. Choi, J. Prado, S. Mangold, and S. Shankar, "IEEE 802.11e contention-based channel access (EDCF) performance evaluation," in *IEEE ICC'03*, May 2003.

[5] 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.

[6] C. H. Foh and M. Zukerman, "Performance analysis of the IEEE 802.11 MAC protocol," in *European Wireless 2002*, Florence, Italy, Feb. 2002.

[7] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service," *IEEE transactions on Wireless Communications*, to appear.

[8] *IEEE standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, ISO/IEC 8802-11: 1999(E), 1999.

[9] *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.

[10] S. Jamin, P.B. Danzig, S. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated service packet networks," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, Feb. 1997.

[11] L. Kleinrock. *Queueing Systems, volume I*. John Wiley & Sons, 1975

[12] Z. Kong, D. H. K. Tsang, and B. Bensaou, "Performance analysis of IEEE 802.11e contention-based channel access," *IEEE JSAC*, Dec. 2004.

[13] OPNET Modeler 10.0. http://www.opnet.com.

[14] A. Veres, A. T. Campbell, M. Barry, and L.-H. Sun, "Supporting service differentiation in wireless packet networks using distributed control," *IEEE JSAC*, Oct. 2001.

[15] 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 *IEEE INFOCOM'02*, New York, June 2002.

[16] Y. Xiao, H. Li, and S. Choi, "Protection and guarantee for voice and video traffic in IEEE 802.11e Wireless LANs," in *IEEE INFOCOM'04*, March 2004.

[17] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *Journal of Wireless Communications and Mobile Computing*, vol. 4, p 917-931, Dec. 2004.

[18] H. Zhai, X. Chen, and Y. Fang, "A call admission and rate control scheme for multimedia support over IEEE 802.11 wireless LANs," in *QShine'04*, Oct. 2004.

[19] ITU-T G.114. One-way transmission time, 1996.

[20] ITU-T G.1010. End-user multimedia QoS categories, 2001.