

Improving WLAN VoIP Capacity Through Service Differentiation

Deyun Gao, *Member, IEEE*, Jianfei Cai, *Senior Member, IEEE*, Chuan Heng Foh, *Member, IEEE*, Chiew-Tong Lau, *Member, IEEE*, and King Ngi Ngan, *Fellow, IEEE*

Abstract—Voice over Internet protocol (VoIP) is one of the most important applications for IEEE 802.11 wireless local area networks (WLANs). For network planners who deploy VoIP over WLANs, one of the important issues is the VoIP capacity. Directly implementing VoIP over infrastructure WLANs will create the bottleneck problem at the access point (AP). In this paper, we propose the use of the service differentiation provided by the new IEEE 802.11e standard to solve the bottleneck problem and improve voice capacity. In particular, we propose the allocation of a higher priority access category (AC) to the AP while allocating lower priority AC to mobile stations. We develop a simple Markov chain model, which considers the important enhanced distributed channel access (EDCA) parameters and the channel errors under saturation and nonsaturation conditions. Based on the developed analytical model, we analyze the performance of VoIP over EDCA. By appropriately selecting the EDCA parameters, we are able to differentiate the services for the downlink and uplink. The experimental results are very promising. With the adjustment of only one EDCA parameter, we improve the VoIP capacity by 20%–30%.

Index Terms—Enhanced distributed channel access (EDCA), IEEE 802.11e wireless local area networks (WLANs), medium access mechanisms, service differentiation, voice capacity, voice over Internet protocol (VoIP).

I. INTRODUCTION

DUE to the high performance versus price ratio, IEEE 802.11-based wireless local area networks (WLANs) have been massively deployed in public and residential places for various wireless applications. The 802.11 standards include a set of specifications developed by the IEEE for the WLAN technology. In 802.11 WLANs, the medium access control (MAC) layer defines the procedures for 802.11 stations to share a common radio channel, which includes the following two MAC mechanisms: 1) the mandatory distributed coordination function (DCF) and 2) the optional point coordination function

(PCF) [1]. However, the lack of a built-in mechanism for supporting real-time services makes it difficult for the original IEEE 802.11 standard to provide quality of service (QoS) for multimedia applications [2]. Therefore, to enhance QoS support in WLANs, a new standard called IEEE 802.11e [3]–[5] is being developed, which introduces a so-called hybrid coordination function (HCF) for MAC. In particular, the HCF includes the following two medium access mechanisms: 1) contention-based channel access and 2) controlled channel access. The contention-based channel access is referred to as an enhanced distributed channel access (EDCA), which can be regarded as an extension of the DCF, and the controlled channel access is referred to as an HCF controlled channel access (HCCA), which is an extension of the PCF.

While the IEEE 802.11e specification has yet to be accepted as a final standard, the wireless fidelity (Wi-Fi) body, which is the marketing and certification body for WLAN systems, has come out with similar specifications that are very close to the EDCA and HCCA mechanisms. The Wi-Fi versions of the EDCA and HCCA protocols are called Wi-Fi multimedia (WMM) and wireless multimedia scheduled access (WMM-SA), respectively [6]. On the other hand, many real-time multimedia applications have also been developed and running over WLANs. Among the various applications, voice over Internet protocol (VoIP) is recognized as one of the most important applications for WLANs. However, for the deployment of VoIP over WLANs, many challenges still remain, including QoS, call admission control, network capacity, etc.

Recently, quite a few research work related to the problem of the WLAN VoIP capacity has appeared in the literature. Here, the WLAN VoIP capacity is defined as the maximum number of voice connections supported in WLANs. VoIP capacity has been evaluated through either simulation or experimental testbeds in [7]–[9]. There are also some papers that propose the utilization of the 802.11 MAC mechanisms to improve the voice capacity. In particular, in [10], Hwang and Cho proposed a new access scheme for VoIP packets in 802.11e WLANs, where the access point (AP) can transmit its VoIP packets after a PCF interframe space without backoff or contention. In [11], Gopalakrishnan *et al.* proposed to aggregate voice packets at the AP and use the elements in the 802.11 fragmenting procedure to transmit them. In [12], Wang *et al.* designed a voice multiplex–multicast scheme, which eliminates inefficiency in the downlink VoIP traffic by multiplexing voice packets from several VoIP streams into one multicast packet. Although all these schemes can improve the WLAN VoIP capacity in one way or another, they require the modifications to the WLAN

Manuscript received March 20, 2006; revised August 11, 2006, March 6, 2007, and April 11, 2007. This work was supported in part by the Science and Engineering Research Council, Singapore Agency for Science, Technology, and Research, under Grant 032 101 0006. The review of this paper was coordinated by Prof. D. O. Wu.

D. Gao is with Beijing Jiaotong University, Beijing 100044, China.

J. Cai, C. H. Foh, and C.-T. Lau are with Nanyang Technological University, Singapore 639798.

K. N. Ngan is with the Chinese University of Hong Kong, Shatin, NT, Hong Kong.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVT.2007.905245

standards. In addition, few papers consider the scenario of VoIP over 802.11e WLANs. Although Shankar *et al.* [13] analyzed the voice capacity in the 802.11a/e WLANs, they only took into account the VoIP traffic transported from mobile stations to the AP and did not consider the downlink VoIP traffic. In fact, the delivery of the downlink VoIP traffic is the primary limiting factor for the VoIP capacity in infrastructure WLANs because for two-way VoIP communications, the AP needs to transmit all the downlink traffic over the radio channel.

In this paper, we propose the use of the service differentiation provided by IEEE 802.11e EDCA [5] to solve the bottleneck problem of VoIP over WLANs and improve voice capacity. In particular, we propose the allocation of higher priority access category (AC) to the AP while allocating lower priority AC to mobile stations. The challenge is how to choose the optimal EDCA parameters so that the maximal VoIP capacity can be achieved. To analyze the performance of EDCA, we propose a Markov chain model under saturation and nonsaturation conditions and also adapt the model to VoIP applications. The experimental results are very promising. With the adjustment of only one EDCA parameter, we improve the VoIP capacity by 20%–30%.

The rest of the paper is organized as follows. In Section II, we give a brief overview of the contention-based mechanisms in 802.11 and 802.11e. Section III describes the problems of VoIP over WLANs. In Section IV, we develop an analytical model for VoIP applications. Based on the analytical model, in Section V, we propose a method to choose the EDCA parameters to improve the WLAN VoIP capacity. Finally, conclusions are drawn in Section VI.

II. OVERVIEW OF CONTENTION-BASED MEDIA ACCESS MECHANISMS

In this section, we briefly introduce and compare the following two contention-based media access mechanisms: 1) the legacy DCF in 802.11 and 2) the latest EDCA in 802.11e.

A. DCF

DCF is based on a carrier sense multiple access/collision avoidance, where stations listen to the medium to determine when it is free. If a station has frames to send and senses that the medium is busy, it will defer its transmission and initiate a backoff counter. The backoff counter is a uniformly distributed random number between zero and the contention window (CW). Once the station detects that the medium has been free for a duration of DCF interframe space (DIFS), it starts a backoff procedure, i.e., decrementing its backoff counter as long as the channel is idle. If the backoff counter has reduced to zero and the medium is still free, the station begins to transmit. If the medium becomes busy in the middle of the decrement, the station freezes its backoff counter and resumes the countdown after deferring for a period of time, which is indicated by the network allocation vector stored in the winning station's frame header.

It is possible that two or more stations begin to transmit at the same time. In such a case, a collision occurs. Collisions

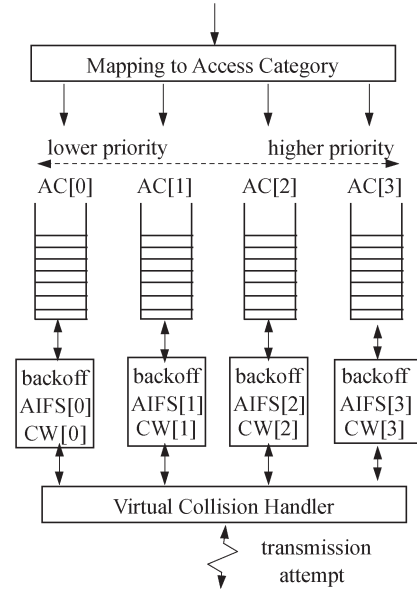


Fig. 1. Virtual backoff of the four ACs.

are inferred by no acknowledgement (ACK) from the receiver. After a collision occurs, all the involved stations double their CWs (up to a maximum value CW_{max}) and compete with the medium again. If a station succeeds in channel access (inferred by the reception of an ACK), the station resets its CW to CW_{min} .

We can see that the DCF does not provide QoS support because all stations operate with the same channel access parameters and have the same medium access priority. There is no mechanism to differentiate different stations and different traffic.

B. EDCA

In the 802.11e standard, the EDCA mechanism extends the DCF access mechanism to enhance the QoS support in the MAC layer by introducing multiple ACs to serve different types of traffic. In particular, a station provides four ACs that have independent transmission queues, as shown in Fig. 1. Each AC, which is basically an enhanced variant of the DCF, contends for a transmission opportunity (TXOP) using one set of the EDCA channel access parameters. A TXOP is an interval of time when a particular QoS station (QSTA) has the right to initiate frame exchange sequences onto the wireless medium. The TXOP is either obtained by the QSTA by successfully contending for the channel or assigned by the hybrid coordinator. If a TXOP is obtained using the contention-based channel access, it is defined as an EDCA TXOP. In this paper, we just use TXOP to represent EDCA TXOP for simplicity. The set of the EDCA channel access parameters include the following.

- 1) $CW_{min}[AC]$: minimal CW value for a given AC. CW_{min} can be different for different ACs. Assigning smaller values of CW_{min} to high-priority classes can ensure that high-priority classes can obtain more TXOPs than low-priority ones.

- 2) $CW_{\max}[\text{AC}]$: maximal CW value for a given AC. Similar to CW_{\min} , CW_{\max} is also on a per AC basis.
- 3) $\text{AIFS}[\text{AC}]$: arbitration interframe space. Each AC starts its backoff procedure after the channel is idle for a period of $\text{AIFS}[\text{AC}]$ instead of DIFS.
- 4) $\text{TXOP}_{\text{limit}}[\text{AC}]$: the limit of consecutive transmission. During a TXOP, a station is allowed to transmit multiple data frames but is limited by $\text{TXOP}_{\text{limit}}[\text{AC}]$.

Note that if the backoff counters of two or more ACs collocated in the same station elapse at the same time, a scheduler inside the station treats the event as a “virtual collision.” The TXOP is given to the AC with the highest priority among the colliding ACs, and the other colliding ACs would defer and try again later as if the collision occurred in the real medium. More details can be found in [5]. In short, through service differentiation among multiclass traffic, EDCA provides a very powerful platform to support QoS in WLANs for multimedia applications.

III. PROBLEMS WITH VOIP OVER WLANS

Addressing the challenges of running VoIP over WLANs requires an understanding of user expectations, technology requirements for telephony, and basic WLAN operations.

There are only a few voice codec standards used for the IP telephony. Typically, voice is coded with the G.711 codec at a rate of 64 kb/s, which is further divided into raw voice packets. Before these application-layer raw voice packets arrive at the MAC layer, they are expanded with some protocol headers, including 12 B for real-time transport protocol, 8 B for user datagram protocol, and 20 B for IP. Considering a codec packetization interval of 20 ms, the raw voice packet is 160 B. From the viewpoint of the MAC layer, the frame payload size is $160 + 40 = 200$ B, and the data rate is $200 \times 8/20 = 80$ kb/s.

We consider a common scenario of VoIP over WLAN, as shown in Fig. 2, where an AP and many mobile stations form a single 802.11 basic service set (BSS). The BSS is connected to the Internet via the AP. A voice talk typically involves one WLAN mobile user and another user connected to Internet. For simplicity, in this paper, we only consider the wireless link and ignore the impact of wired links if the communication path includes both wireless and wired links.

It is well known that a VoIP connection has a few QoS requirements, including the throughput requirement, the 150-ms end-to-end delay requirement, and the requirements of low delay jitter and effectively 0% packet loss. Garg and Kappes [7] show that the packet loss rate of VoIP traffic due to the overflow of the queue is very small in the cases where the system capacity is not exceeded and the number of allowed retransmissions is large enough (e.g., using the default retry limit of seven). This claim is also supported in an independent work by Hole and Tobagi [9], where they show that the access delay is also low in WLANs under the similar conditions. Therefore, in this paper, we focus on the throughput requirement for the performance of VoIP connections.

Ideally, the number of simultaneous VoIP sessions that can be supported by the IEEE 802.11b WLAN is around $11 \text{ Mb/s} / (2 \times$

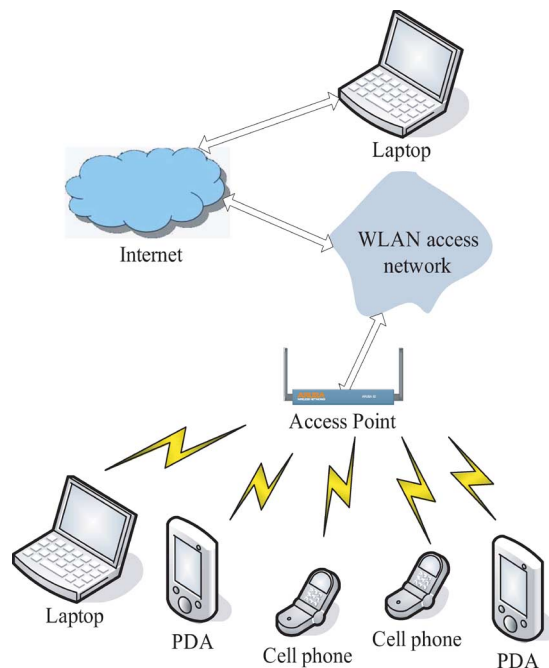


Fig. 2. Common scenario of VoIP over WLANs.

$80 \text{ kb/s}) = 68$, where a two-way voice communication contains two VoIP streams. However, in practice, the number of VoIP sessions that can be supported by the existing WLANs is much less. This is mainly due to the VoIP connection bottleneck at the AP, resulting in an underutilization of channel bandwidth. In the VoIP application, the AP is often served as the gateway between the local and the remote VoIP clients. As a result, the AP must handle all downlink VoIP traffic in the network. Because EDCA does not differentiate between stations and the AP, the AP equally competes and fairly shares with all the other stations for the channel bandwidth usage. Unaware that the AP requires N times more bandwidth than each VoIP local client to handle N simultaneous VoIP sessions, congestion occurs at the AP before the channel reaches its throughput saturation point.

We conduct a simple study to illustrate the phenomenon of connection bottleneck at the AP for VoIP over WLANs. The considered scenario consists of N VoIP sessions, where each VoIP client in the IEEE 802.11b WLAN communicates with a remote partner via the AP. Each VoIP session produces $2 \times 80 \text{ kb/s}$ of two-way voice communications, where the uplink traffic of each VoIP session is transmitted by each local VoIP client, and the downlink traffic is transmitted by the AP.

In Fig. 3, we plot the throughput of a particular VoIP session, given a number of simultaneous VoIP sessions in the WLAN. The two lines in the figure represent the uplink traffic transmitted by the VoIP client and the downlink traffic transmitted by the AP. These analytical results are obtained from the model detailed in Section IV. They are also validated by the simulation (shown in symbols in Fig. 3). An immediate observation is that the IEEE 802.11b WLAN can adequately accommodate up to only 11 VoIP sessions. Beyond this number, the downlink transmission throughput significantly decreases. This analysis is similar to the previous simulation findings [7]–[9] that, as the number of VoIP sessions increases beyond 11, the channel

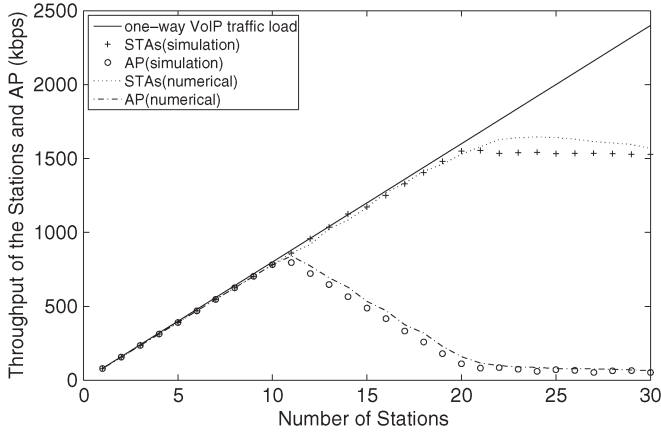


Fig. 3. Simulation results of the throughput for the VoIP over the WLAN.

utilization of the AP drops. This drop in channel utilization at the AP is the result of congestion at the AP because the AP has entered the saturation state.

Interestingly, whereas the AP has become saturated, the channel has yet to reach its saturation point because the uplink transmission throughput does not drop until there are over 20 VoIP sessions. However, beyond 11 VoIP sessions, the AP fails to capture the necessary bandwidth for the VoIP downlink transmissions. Knowing this connection bottleneck phenomenon of VoIP over WLANs, in this paper, we propose a solution that improves the VoIP capacity over WLANs by applying service differentiation between stations and the AP. This allows the AP to capture more bandwidth than each of the stations.

IV. PERFORMANCE ANALYSIS OF EDCA

The QoS in the IEEE 802.11 WLANs is supported using the EDCA specified in the IEEE 802.11e standard. Although the standard defines service differentiation for various traffic types, there has been no service differentiation among communication devices, including stations and APs. Lacking service differentiation for different devices leads to the connection bottleneck at the AP that was discussed in Section III.

In this paper, we propose the extension of EDCA for service differentiation between various device types. To facilitate the design of such an extension, an analytical model is developed to study the influence of the EDCA parameters to the VoIP performance and to identify a set of EDCA parameters that achieves improved voice capacity.

There has been a number of EDCA models [14]–[17], each with a different focus of performance study. However, the majority of the models deal with saturation traffic condition, which is not adequate to describe voice traffic. A modified model that offers nonsaturation traffic description is necessary for our study. Based on Bianchi's original work [18], we enhance it to include an EDCA operation and an error-prone channel consideration.

A. Markov Chain Model

Suppose there are N stations and one AP. We use two ACs AC[up] and AC[dw] for the uplink traffic at the stations and

the downlink traffic at the AP, respectively. For simplicity, we assume $\text{AIFS}[\text{up}] = \text{AIFS}[\text{dw}]$ and do not consider the AIFS differentiation. Fig. 4 shows our proposed simplified Markov chain model. In particular, time is slotted, and each state represents an AC in a particular period. At each state, a state transition is triggered by the occurrence of an event. A state is completely characterized by a three-tuple vector (i, j, k) , where i is the AC index, j denotes the backoff stage, and k denotes the backoff counter. Similar to the approach used in [19] for the nonsaturation extension of Bianchi's Markov chain model, we introduce a new state $(i, -1)$, which indicates that there is no packet awaiting for transmission in the stations, to include nonsaturation traffic consideration. The variable q_i represents the probability that after a successful transmission by a station using AC[i], its queue remains empty after either an idle or a busy slot duration. This input variable provides the nonsaturation load adjustment of a station. Setting this probability to zero reduces the Markov chain model to that of the saturation load condition.

We assume that $P_{i,f}$, which is the unsuccessful transmission probability of AC[i], and $P_{i,b}$, which is the channel busy probability observed by the AC[i] queue, are constant and independent of the backoff procedure. Unlike the previous model [14], the probability of $P_{i,f}$ in our proposed model consists of the following two parts: 1) the collision probability P_i and 2) the failed transmission probability P_e due to transmission errors. Mathematically, $P_{i,f}$ can be expressed as

$$P_{i,f} = 1 - (1 - P_i)(1 - P_e) = P_i + P_e - P_i P_e \quad (1)$$

and P_e is calculated by

$$P_e = 1 - (1 - \epsilon)^l \quad (2)$$

where ϵ is the channel bit error rate, and l is the frame length in bits. Note that, here, we assume that the collision probability P_i and the packet error rate (PER) P_e are independent and that the bit errors are memoryless. Setting P_e to zero reduces the Markov chain model to that of the perfect channel conditions.

Here, $W_{i,j}$ is the length of the CW for AC[i] at backoff stage j , and m_i and h_i denote the maximum number of retransmission using different $W_{i,j}$ and the maximum $W_{i,j}$, respectively. For a different backoff stage j ($0 \leq j \leq m_i + h_i$), the length of the corresponding CW is given by

$$W_{i,j} = \min [CW_{\max}[i] + 1, 2^j (CW_{\min}[i] + 1)] \quad (3)$$

where $CW_{\max}[i] + 1 = 2^{m_i} (CW_{\min}[i] + 1)$, and $W_{i,0} = W_i$.

Let $b_{i,j,k}$ denote the stationary probability for the state $\{i, j, k\}$. According to the regularity of the Markov chain, we have the following relationships:

$$b_{i,j-1,0} P_{i,f} = b_{i,j,0} \quad (4)$$

$$b_{i,-1} = \frac{q}{1-q} b_{i,0,0} \quad (5)$$

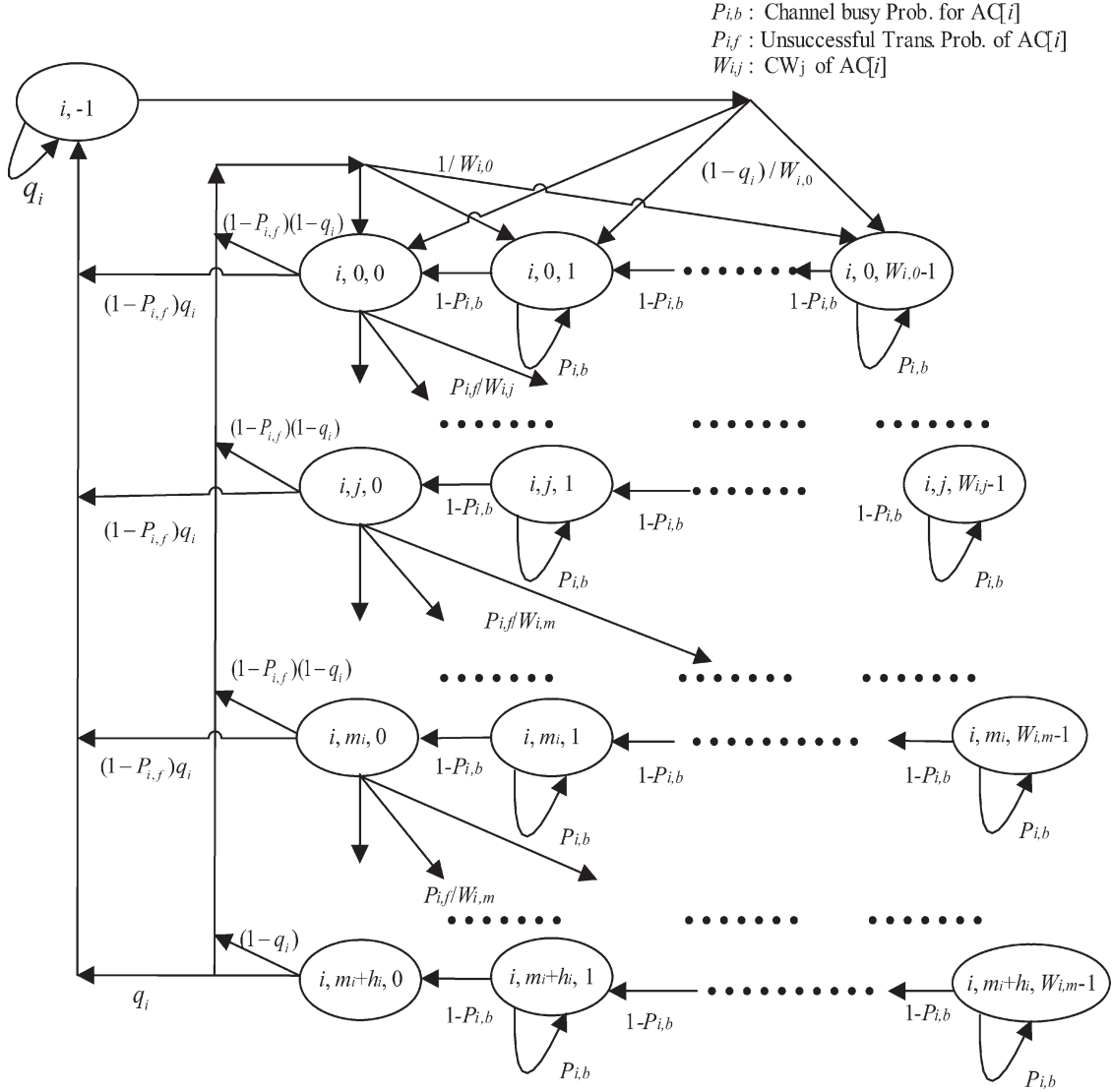


Fig. 4. Transition diagram of the proposed Markov chain model for AC[i].

and

$$b_{i,j,k} = \frac{W_{i,j} - k}{(1 - P_{i,b})W_{i,j}} b_{i,j,0} \quad (6)$$

$j \in [0, m_i + h_i]; \quad k \in [1, W_{i,j} - 1].$

This way, all the values of $b_{i,j,k}$ can be expressed in terms of $P_{i,b}$, $P_{i,f}$, and $b_{i,0,0}$. Because the summation of all the state probabilities should be equal to one, we obtain

$$b_{i,0,0} = \left[\frac{q_i}{1 - q_i} + \frac{A}{2(1 - P_{i,b})(1 - 2P_{i,f})(1 - P_{i,f})} \right]^{-1} \quad (7)$$

where

$$\begin{aligned}
 A = & (1 - 2P_{i,b})(1 - 2P_{i,f}) \left(1 - P_{i,f}^{m_i + h_i + 1} \right) \\
 & + W_i \left[1 - (2P_{i,f})^{m_i + 1} \right] (1 - P_{i,f}) \\
 & + W_i (2P_{i,f})^{m_i} P_{i,f} \left(1 - P_{i,f}^{h_i + 1} \right) (1 - 2P_{i,f}).
 \end{aligned}$$

From (7), we can see that $b_{i,0,0}$ is determined by $P_{i,b}$, $P_{i,f}$, and q_i . Now, the problem is how to calculate $P_{i,b}$, $P_{i,f}$, and q_i .

Let us first consider the probability τ_i that one AC[i] tries to access the medium. It is clear that τ_i should be equal to all the steady-state probabilities of the states $\{i, j, 0\}$, $j = 0, 1, \dots, m_i + h_i$, where the backoff counter reaches zero. That is

$$\tau_i = \sum_{j=0}^{m_i + h_i} b_{i,j,0} = \frac{1 - P_{i,f}^{m_i + h_i + 1}}{1 - P_{i,f}} b_{i,0,0}. \quad (8)$$

Obviously, the channel is deemed idle by one station if the other stations and the AP do not use it; otherwise, it is sensed busy. Similarly, the channel is deemed idle by the AP if no station uses it. As a result, $P_{i,b}$, which is the probability that the channel is observed busy by AC[i], can be derived as

$$P_{i,b} = \begin{cases} 1 - (1 - \tau_{\text{up}})^{N-1} (1 - \tau_{\text{dw}}), & i = \text{up} \\ 1 - (1 - \tau_{\text{up}})^N, & i = \text{dw}. \end{cases} \quad (9)$$

As for calculating $P_{i,f}$ defined in (1), we need to compute the collision probability P_i . Clearly, P_i is equal to $P_{i,b}$.

B. Throughput

We are interested in the uplink and downlink throughput, which is denoted as $S_i(N)$ ($i \in \{\text{up}, \text{dw}\}$). Because each station and the AP operate according to the state transition diagram shown in Fig. 4, we consider the period that all the ACs remain in their states as a time interval. This way, $S_i(N)$ is calculated according to the ratio of the time occupied by the AC $[i]$'s delivered packet to the average length of the time interval, i.e.,

$$\begin{aligned} S_i(N) &= R \frac{E[\text{time for successful transmission in an interval}]}{E[\text{length between two consecutive transmissions}]} \\ &= R \frac{P_{i,s}E[P]}{E[I] + E[NC] + E[C]} \end{aligned} \quad (10)$$

where R is the bandwidth of the WLAN, $E[P]$ is the VoIP payload length, $P_{i,s}E[P]$ is the average amount of successfully transmitted payload information, and the average length of a time interval consists of the following three parts: 1) $E[I]$, which is the expected value of idle time before a transmission; 2) $E[NC]$, which is the transmission time without collision; and 3) $E[C]$, which is the collision time.

The successful transmission probability $P_{i,s}$ of the station and the AP can be calculated as

$$P_{i,s} = \begin{cases} \frac{\tau_{\text{up}}(1-\tau_{\text{up}})^{N-1}(1-\tau_{\text{dw}})}{1-P_b}(1-P_e), & i = \text{up} \\ \frac{\tau_{\text{dw}}(1-\tau_{\text{up}})^N}{1-P_b}(1-P_e), & i = \text{dw} \end{cases} \quad (11)$$

where P_b , which is the channel busy probability, is defined as

$$P_b = 1 - (1 - \tau_{\text{up}})^N(1 - \tau_{\text{dw}}). \quad (12)$$

Note that P_b is different from $P_{i,b}$ in (9). $P_{i,b}$ is the channel busy probability observed by one AC $[i]$, whereas P_b is the channel busy probability from the network point of view.

The expected value of the idle time in a time interval $E[I]$ can be easily estimated as $[(1 - P_b)/P_b]\sigma$, where σ is the value of one system slot. The transmission time without collision $E[NC]$ includes the following two parts: 1) the successful transmission time $E[S]$ and 2) the failure transmission time only due to transmission errors $E[TE]$. We can derive $E[S]$ and $E[TE]$ as

$$\begin{aligned} E[S] &= (NP_{\text{up},s} + P_{\text{dw},s})(T_s + \text{AIFS}[\text{AC}]) \\ E[TE] &= (NP_{\text{up},e} + P_{\text{dw},e})(T_e + \text{AIFS}[\text{AC}]) \end{aligned} \quad (13)$$

where T_s and T_e are the average time of a successful transmission, the average times of a noncollision failure transmission

$P_{\text{up},s}$ and $P_{\text{dw},s}$ are given in (11), and the probabilities of a noncollision failure transmission $P_{\text{up},e}$ and $P_{\text{dw},e}$ are given by

$$P_{i,e} = \begin{cases} \frac{\tau_{\text{up}}(1-\tau_{\text{up}})^{N-1}(1-\tau_{\text{dw}})}{1-P_b}P_e, & i = \text{up} \\ \frac{\tau_{\text{dw}}(1-\tau_{\text{up}})^N}{1-P_b}P_e, & i = \text{dw}. \end{cases} \quad (14)$$

The collision time $E[C]$ can be expressed as

$$E[C] = P_c T_c \quad (15)$$

where T_c is the average collision time, P_c is the collision probability, and $P_c = 1 - NP_{\text{up},s} - P_{\text{dw},s} - NP_{\text{up},e} - P_{\text{dw},e}$.

The values of T_s , T_e , and T_c in (13) and (15) depend on the channel access mode. In the 802.11e standard, the channel access mode is more complicated than that of the legacy 802.11 standard. In particular, in addition to the basic access mode and the request-to-send/clear-to-send access mode, the 802.11e standard also introduces other mechanisms such as no ACK and block ACK. In this paper, we only consider the basic access mode. For other access modes, the values of T_s , T_e , and T_c can be obtained by applying the same procedure. For the basic access mode, we have

$$\begin{cases} T_s = \text{AIFS} + H + E[P] + \text{SIFS} + \delta + \text{ACK} + \delta \\ T_c = \text{AIFS} + H + E[P] + \delta \end{cases}$$

where ACK is the time duration of an ACK frame, $H = \text{PHY}_{\text{hdr}} + \text{MAC}_{\text{hdr}}$ is the header duration, δ is the propagation delay, and the payload length $E[P]$ is the packet length of voice packets. As for T_e , we assume that transmission errors do not occur in the duration announcements (contained in the preamble/header part of a frame), and thus, $T_e = T_s$. For more general cases of transmission error distribution, the calculation of T_e can be found in [20].

C. VoIP Capacity

We first revisit the illustration of the connection bottleneck at the AP given in Fig. 3. The scenario considers one AP with N other stations implementing the IEEE 802.11b MAC protocol. We investigate the VoIP capacity using our developed model with the IEEE 802.11b parameter set. To study the maximum capacity, we consider an error-free channel. Define $U_{\text{up}}(N)$ and $U_{\text{dw}}(N)$ to be the throughput (in kilobits per second) of a particular VoIP session in the uplink and downlink transmissions, respectively. Given that each VoIP session produces 2×80 kb/s, the numerical computation for $U_{\text{dw}}(N)$ requires that

$$U_{\text{dw}}(N) = \begin{cases} 80, & (\exists q_{\text{dw}})(S_{\text{dw}}(N) = 80N) \\ \frac{S_{\text{dw}}(N)}{N}, & \text{otherwise} \end{cases} \quad (16)$$

where the second condition in the aforementioned expression describes the saturation of the AP. Similarly, for $U_{\text{up}}(N)$, we use

$$U_{\text{up}}(N) = \begin{cases} 80, & (\exists q_{\text{up}})[S_{\text{up}}(N) = 80] \\ S_{\text{up}}(N), & \text{otherwise.} \end{cases} \quad (17)$$

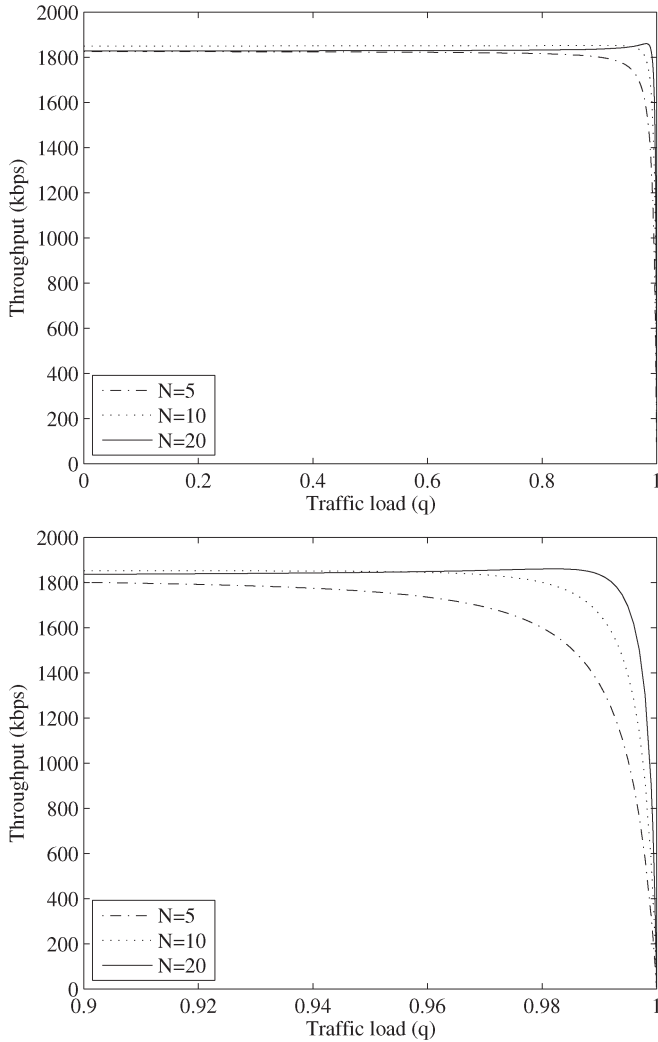


Fig. 5. Throughput of the WLAN with the traffic load. (a) $0 \leq q_{\text{up}} \leq 1$. (b) $0.9 \leq q_{\text{up}} \leq 1$.

As can be seen in Fig. 3, the connection bottleneck at the AP appears when there are more than 11 VoIP simultaneous sessions in the IEEE 802.11b WLAN. The consequence of the connection bottleneck at the AP is the underutilization of the channel bandwidth.

To fully utilize the channel bandwidth, it is necessary to evaluate the maximum achievable capacity in the IEEE 802.11 WLAN. We consider a scenario of N stations (with no AP), with each station gradually increasing its traffic load toward its saturation. In the analytical model, this is done by increasing the traffic load by adjusting the quantity q_{up} . In the extreme case, where $q_{\text{up}} = 0$, the model is reduced to the case of that saturated the traffic load condition. We depict the channel throughput versus the traffic load in Fig. 5 for several N values. As can be seen, while the maximum achievable throughput of the WLAN does not fall at the point of the saturated load (i.e., $q_{\text{up}} = 0$), the saturation throughput does not differ much from the maximum achievable throughput. This suggests that the saturation throughput can be used to describe the maximum achievable throughput for our studied problem.

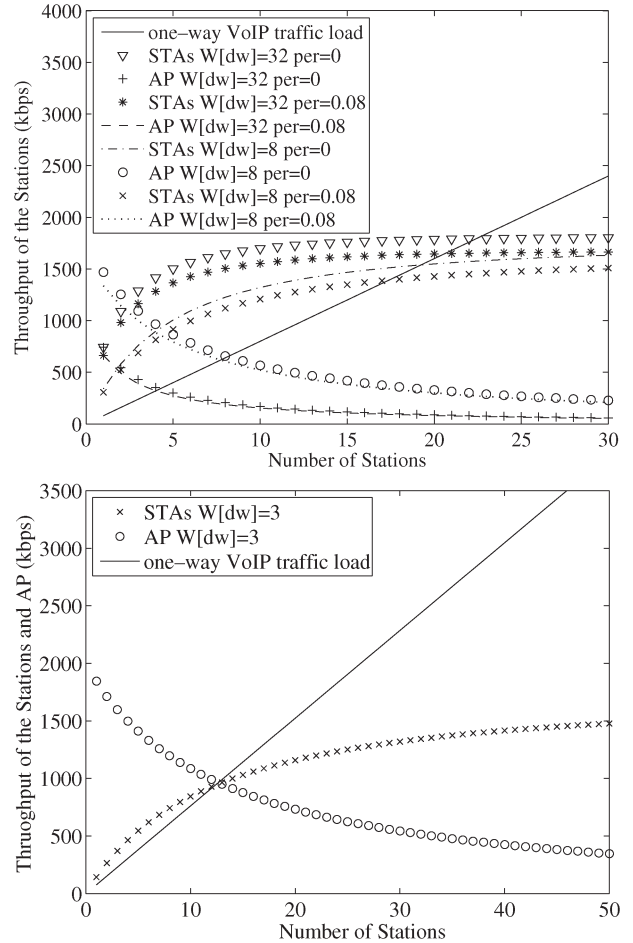


Fig. 6. Saturation throughput of the stations and the AP. (a) For different W_{dw} and different PERs. (b) For $W_{\text{dw}} = 3$ and PER = 0.

V. SELECTING THE EDCA PARAMETERS

As mentioned at the beginning of Section IV, to overcome the bottleneck problem of VoIP, we propose the application of EDCA to provide different services to the AP and the stations. The key challenge is the selection of the optimal EDCA parameters. Considering the symmetric property of VoIP traffic and the one-way throughput requirement R_{req} , the operating points should be those at which the downlink throughput is equal to the aggregated uplink throughput, as well as to NR_{req} . As justified in Section IV, because saturation throughput well represents the maximum achievable throughput of the WLAN, our succeeding study uses the saturation case, i.e., setting $q_i = 0$, of the EDCA described in Section IV to derive the EDCA parameters. Although our obtained solution may not be the optimal one, it achieves the improvement of the WLAN VoIP capacity. For simplicity, we only consider adjusting one EDCA parameter, i.e., W_{dw} . Other EDCA parameters can be changed in a similar way.

Fig. 6(a) shows the numerical throughput results for the uplink and the downlink under different values of W_{dw} . It can be observed that adjusting the CW_{min} of the AP enables a tradeoff between the uplink and downlink saturation throughput. Note that, in Fig. 6, when the number of stations is small,

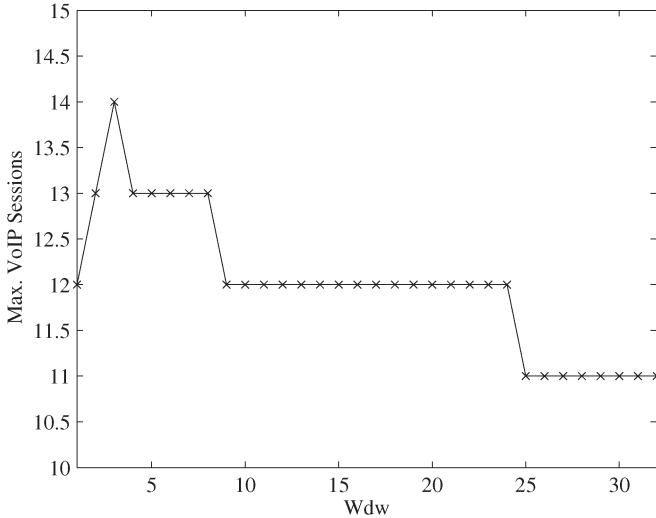


Fig. 7. Maximum VoIP sessions with the parameter W_{dw} .

the throughput is larger than the input traffic load, which is not realistic. This is because the throughput that we plot is the saturation throughput, while the cases of small numbers of stations are actually under unsaturation conditions. We notice that, for the same PER, when we reduce the W_{dw} , the throughput of the AP is increased and becomes increasingly close to the aggregated uplink throughput. In particular, the cross point between the aggregated uplink and downlink throughput moves to the upper right. By continuously decreasing W_{dw} , we can make the cross point meet the one-way traffic line, as shown in Fig. 6(b). Clearly, this is a good operational point, where $S_{dw}(N) = NS_{up}(N) = NR_{req}$. Although at this point, the AP and all the stations might still be under unsaturation conditions, we know that the satisfaction of the throughput requirement can be guaranteed. If we further move the cross point, both the downlink and the aggregated uplink saturation throughput become less than NR_{req} , although the actual throughput might not be. Therefore, with the saturation model, the best solution that we can obtain is the one shown in Fig. 6(b), where with $W_{dw} = 3$ and $PER = 0$, we achieve an improved voice capacity of 14, which corresponds to the $(14 - 11)/11 = 27.3\%$ increase of voice capacity.

In Fig. 7, we give out the maximum VoIP sessions when we change W_{dw} from 1 to 32, which illustrates that the maximum VoIP sessions supported in the WLANs is changed under the different values of W_{dw} , and 3 is the optimal W_{dw} value. Combining with other EDCA parameters, an even larger voice capacity can be achieved. Our solution can be regarded as a lower bound of the maximal voice capacity.

By considering the $W_{dw} = 3$ setting for the AP, we repeat the VoIP capacity analysis using (16) and (17). The numerical results shown in Fig. 8 confirms that the VoIP capacity, indeed, increases to 14 VoIP sessions. The fact that both the throughput of the uplink and downlink traffic drop at the same point also suggests the elimination of the connection bottleneck at either direction of transmissions, as both the AP and the stations become saturated only at the same point when the WLAN is overloaded.

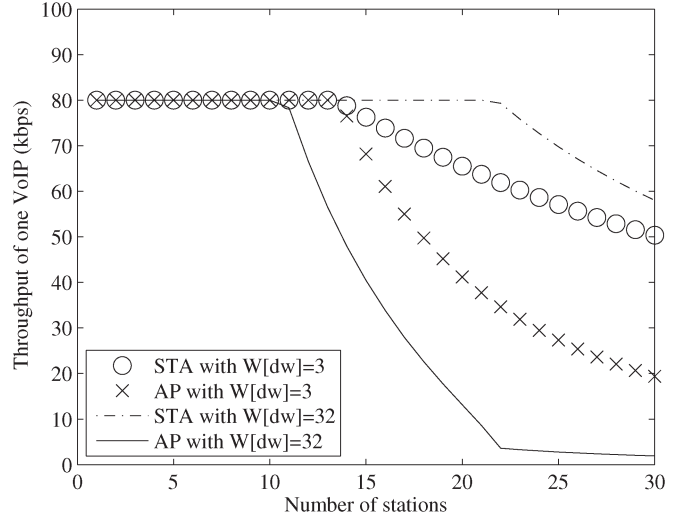


Fig. 8. Throughput of one VoIP session at the stations and the AP with $W_{dw} = 3$ and $W_{dw} = 32$.

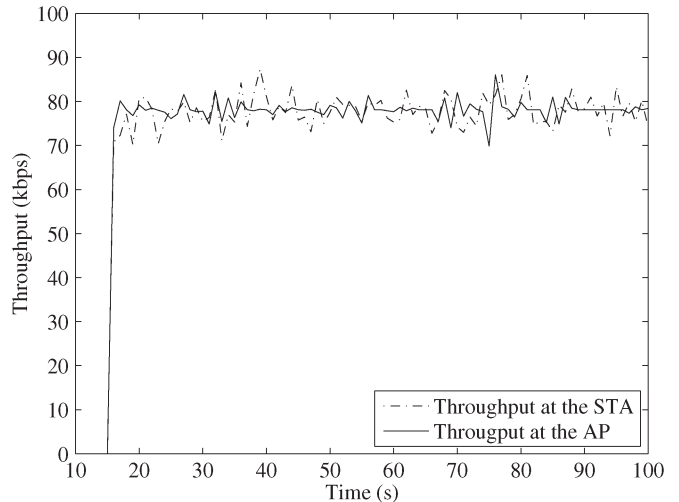


Fig. 9. Throughput of one VoIP session at the stations and the AP.

We further conduct the Network Simulator 2 (NS-2) simulation to measure the QoS performance of the VoIP sessions under the operational point ($W_{dw} = 3$ and $N = 14$) that is determined by the theoretical analysis. Fig. 9 shows the throughput of one VoIP session at the stations and the AP, which is calculated every 1 s. Although the throughput has a little oscillation around 80 kb/s, it basically satisfies the requirement of the VoIP sessions. Fig. 10 shows the delay of one VoIP session at the stations and the AP. Note that the delay includes both access and queuing delays. Fig. 11 plots the corresponding cumulative distribution function of the delay. We can see that most voice packets experience a very small delay. Over 90% of the voice packets have a delay less than 50 ms, and over 95% of the voice packets have a delay less than 75 ms. Therefore, we can conclude that our obtained operational point is a feasible point, which can satisfy the VoIP QoS requirements.

VI. CONCLUSION

In this paper, we have shown that, for VoIP over infrastructure WLANs, the AP is the bottleneck that limits the VoIP

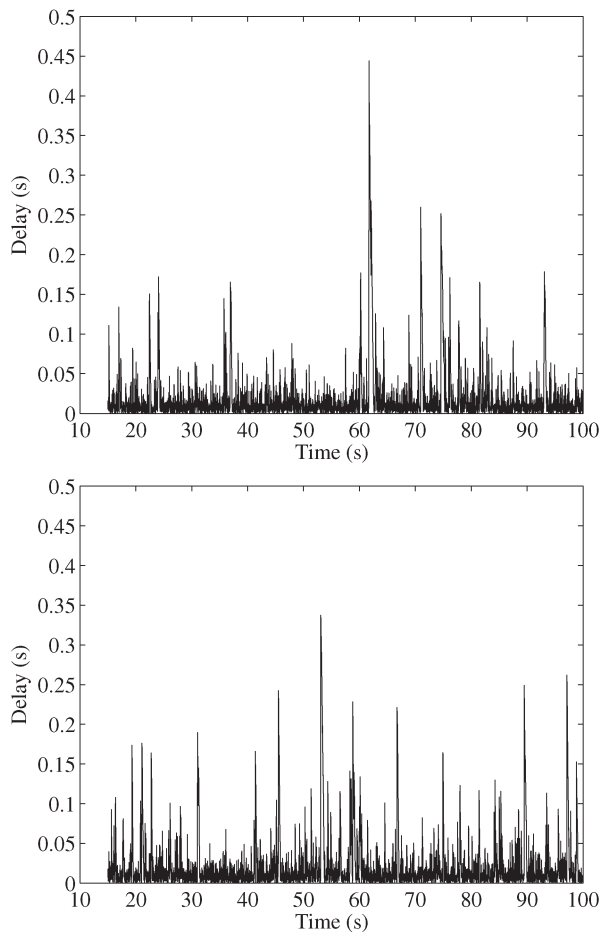


Fig. 10. Delay of one VoIP session. (a) At a station. (b) At the AP.

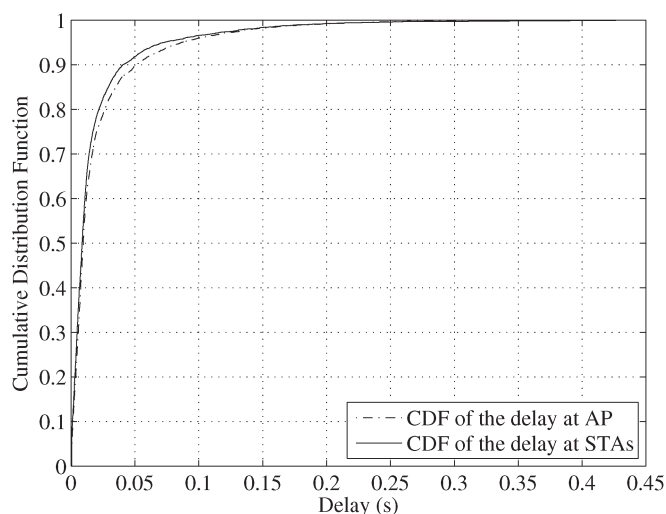


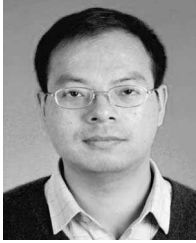
Fig. 11. Cumulative distribution function of the VoIP delay at the stations and the AP.

capacity. We have proposed the use of the 802.11e EDCA mechanism to provide service differentiation to the AP and the mobile stations to improve the WLAN VoIP capacity. In particular, we have developed the Markov chain model for the EDCA performance analysis, which considers the important

EDCA parameters and the channel errors under saturation and nonsaturation conditions. The analytical performance on the saturation throughput for multiclass traffic has been validated via the NS-2 simulations. We have further proposed the bypassing of the unsaturation performance analysis and the use of the developed saturation model to choose the EDCA parameters. Our analytical results have demonstrated that by adjusting only one EDCA parameter, i.e., W_{dw} , our proposed method improves the VoIP capacity by 20%–30%. The NS-2 simulation results have further proved that the analytically selected EDCA parameters can satisfy the VoIP QoS requirements.

REFERENCES

- [1] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std. 802.11, 1999.
- [2] A. Lindgren, A. Almquist, and O. Schelen, "Evaluation of quality of service schemes for IEEE 802.11 wireless LANs," in *Proc. 26th Annu. IEEE Conf. Local Comput. Netw.*, Tampa, FL, Nov. 2001, pp. 348–351.
- [3] X. Yang, "IEEE 802.11e: QoS provisioning at the MAC layer," *Wireless Commun.*, vol. 11, no. 3, pp. 72–79, Jun. 2004.
- [4] S. Mangold, S. Choi, G. Hiertz, O. Klein, and B. Walke, "Analysis of IEEE 802.11e for QoS support in wireless LANs," *Wireless Commun.*, vol. 10, no. 6, pp. 40–50, Dec. 2003.
- [5] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Std. 802.11e-2005, 2005.
- [6] *WMM: Wi-Fi Multimedia*. [Online]. Available: <http://www.wi-fi.org/opensection/wmm.asp>
- [7] S. Garg and M. Kappes, "Can I add a VoIP call?" in *Proc. IEEE ICC*, Anchorage, AK, May 2003, vol. 2, pp. 779–783.
- [8] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g wireless LANs," in *Proc. IEEE GLOBECOM*, Dallas, TX, Dec. 2004, vol. 3, pp. 1549–1553.
- [9] D. Hole and F. Tobagi, "Capacity of an IEEE 802.11b wireless LAN supporting VoIP," in *Proc. IEEE ICC*, Paris, France, Jun. 2004, vol. 1, pp. 196–201.
- [10] G.-H. Hwang and D.-H. Cho, "New access scheme for VoIP packets in IEEE 802.11e wireless LANs," *IEEE Commun. Lett.*, vol. 9, no. 7, pp. 667–669, Jul. 2005.
- [11] P. Gopalakrishnan, D. Famolari, and T. Kodama, "Improving WLAN voice capacity through dynamic priority access," in *Proc. IEEE GLOBECOM*, Paris, France, Jun. 2004, vol. 5, pp. 3245–3249.
- [12] W. Wang, S. C. Liew, and V. Li, "Solutions to performance problems in VoIP over a 802.11 wireless LAN," *IEEE Trans. Veh. Technol.*, vol. 54, no. 1, pp. 366–384, Jan. 2005.
- [13] S. Shankar, J. del Prado Pavon, and P. Wienert, "Optimal packing of VoIP calls in an IEEE 802.11 a/e WLAN in the presence of QoS constraints and channel errors," in *Proc. IEEE GLOBECOM*, Paris, France, Jun. 2004, vol. 5, pp. 2974–2980.
- [14] Z. Kong, D. Tsang, B. Bensaou, and D. Gao, "Performance analysis of IEEE 802.11e contention-based channel access," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 10, pp. 2095–2106, Dec. 2004.
- [15] J. Robinson and T. Randhawa, "Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 5, pp. 917–928, Jun. 2004.
- [16] X. Yang, "Enhanced DCF of IEEE 802.11e to support QoS," in *Proc. IEEE Wireless Commun. Netw. Conf.*, New Orleans, LA, Mar. 2003, vol. 2, pp. 1291–1296.
- [17] J. Tantra, C. H. Foh, and A. Mnaouer, "Throughput and delay analysis of the IEEE 802.11e EDCA saturation," in *Proc. IEEE ICC*, Seoul, Korea, 2005, vol. 5, pp. 3450–3454.
- [18] 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.
- [19] B. Li and R. Battiti, "Analysis of the IEEE 802.11 DCF with service differentiation support in non-saturation conditions," in *Proc. 5th Int. Workshop QoSIS*, Barcelona, Spain, 2004, pp. 64–73.
- [20] X. Dong and P. Varaiya, "Saturation throughput analysis of IEEE 802.11 wireless LANs for a lossy channel," *IEEE Commun. Lett.*, vol. 9, no. 2, pp. 100–102, Feb. 2005.



Deyun Gao (M'06) received the B.Eng. and M.Eng. degrees in electrical engineering and the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 1994, 1999, and 2002, respectively.

He was a Research Associate with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong, and a Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. Since 2007, he has been an

Associate Professor with the School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China. His research interests are in the area of mobile and wireless Internet with emphasis on quality of service guarantee, multimedia traffic delivery, MAC protocol, and next-generation networks.



Jianfei Cai (S'98–M'02–SM'07) received the Ph.D. degree from the University of Missouri-Columbia in 2002.

He is currently an Assistant Professor with Nanyang Technological University, Singapore. His major research interests include digital media processing, multimedia compression, communications, and networking technologies. He is the author of more than 50 technical papers published in international conferences and journals.

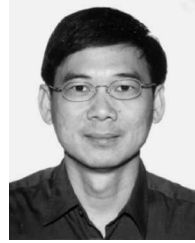
Dr. Cai has actively participated on the program committees of various conferences. He is the Mobile Multimedia Track Cochair for the IEEE International Conference on Multimedia and Expo 2006, the Technical Program Cochair for the International Multimedia Modeling 2007, and the Conference Cochair for Multimedia on Mobile Devices 2007. He is also an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Chuan Heng Foh (S'00–M'03) received the B.S. degree in electronic engineering from Fu Jen Catholic University, Hsien, Taiwan, R.O.C., in 1992, the M.S. degree from Monash University, Clayton, Australia, in 1999, and the Ph.D. degree from the University of Melbourne, Melbourne, Australia, in 2002.

From July 2002 to December 2002, he was a Lecturer with Monash University. He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include protocol

design and performance analysis of mobile wireless and optical networks.



Chiew-Tong Lau (M'90) received the B.Eng. degree from Lakehead University, Thunder Bay, ON, Canada, in 1983 and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, in 1985 and 1990, respectively.

He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His main research interests are in wireless communications.



King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from Loughborough University, Leicestershire, U.K.

He was with Nanyang Technological University, Singapore, and the University of Western Australia, Perth, as a Full Professor. He is currently a Chair Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, NT, Hong Kong. He has served as an Associate Editor for the *EURopean Association for Signal Processing (EURASIP) Journal on Applied Signal Processing*.

He is an Associate Editor for the *Journal on Visual Communication and Image Representation* and an Area Editor for the *EURASIP Journal of Signal Processing: Image Communication*. He has published three authored books, five edited volumes, and over 200 refereed technical papers in the areas of image/video coding and communications.

Dr. Ngan is a Fellow of the Institution of Engineering and Technology (U.K.) and the Institution of Engineers Australia. He has served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has chaired a number of prestigious international conferences on video signal processing and communications and served on the advisory and technical committees of numerous professional organizations. He will cochair the IEEE International Conference on Image Processing that will be held in Hong Kong in 2010.