Improving WLAN VoIP Capacity through Service Differentiation

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

Abstract—Voice over Internet protocol (VoIP) is one of the most important applications for the IEEE 802.11 wireless local area networks (WLANs). For network planners who are deploying VoIP over WLANs, one of the important issues is the VoIP capacity. Directly implementing VoIP over infrastructure WLANs will have the bottleneck problem in the access point (AP). In this paper, we propose to use the service differentiation provided by the new IEEE 802.11e standard to solve the bottleneck problem and improve the voice capacity. In particular, we propose to allocate 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 EDCA parameters, to differentiate the services for the downlink and the 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—VoIP, voice capacity, service differentiation, EDCA, IEEE 802.11e WLANs, medium access mechanisms.

I. INTRODUCTION

Due to the high performance vs price ratio, the 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. It includes two MAC mechanisms: the mandatory distributed coordination function (DCF) and the optional point coordination function (PCF) [1]. However, the lack of a built-in mechanism for supporting real-time services makes the original IEEE 802.11 standard very difficult to provide QoS for multimedia applications [2]. Therefore, in order to enhance the QoS support in WLANs, a new standard called IEEE 802.11e [3], [4], [5] is being developed, which introduces a so-called hybrid coordination function (HCF) for medium access control. In particular, HCF includes two medium access mechanisms: contention-based channel access and controlled channel access. The contention-based channel access is referred as enhanced distributed channel access (EDCA), which can be regarded as an extension of DCF, and the controlled channel access is referred as HCF controlled channel access (HCCA), which is an extension of PCF.

While the IEEE 802.11e specification has yet to be accepted as a final standard, the WiFi body, which is the marketing and certification body for WLAN systems, has come out similar specifications very close to the EDCA and HCCA mechanisms. The WiFi versions of the EDCA protocol and the HCCA protocol are called WiFi 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 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 are still remaining including quality-of-service (QoS), call admission control, and network capacity etc.

Recently, quite a few research work related to the problem of the WLAN VoIP capacity has appeared in literature. Here, the WLAN VoIP capacity is defined as the maximum number of voice connections supported in WLANs. The VoIP capacity has been evaluated through either simulation or experimental testbeds in [7], [8], [9]. There are also some papers proposing to utilize the 802.11 MAC mechanisms to improve the voice capacity. In particular, in [10], the authors proposed a new access scheme that enhances the VoIP capacity in infrastructure WLANs, where the AP can transmit its VoIP packets after PCF interframe space (PIFS) without backoff or contending. 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 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 the other, they require the modifications to the WLAN standards. In addition, few paper considers the scenario of VoIP over 802.11e WLANs. Although the authors in [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 to use the service differentiation provided by the IEEE 802.11e EDCA [5] to solve the bottleneck problem of VoIP over WLANs and improve the voice capacity. In particular, we propose to allocate 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. In order to analyze the performance of EDCA, we propose a Markov chain model under saturation and non-saturation 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 the 802.11 and the 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 two contention-based media access mechanisms: the legacy DCF in 802.11 and the latest EDCA in 802.11e.

A. Distributed Coordination Function

DCF is based on CSMA/CA (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 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 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 back-off 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 (NAV) 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 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, CWmax) and compete the medium again. If a station succeeds in channel access (inferred by the reception of an ACK), the station resets its CW to CWmin.

We can see that DCF does not provide QoS supports since 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. Enhanced Distributed Channel Access

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

- $CW_{\text{min}}[AC]$: minimal CW value for a given AC. CWmin can be different for different ACs. Assigning smaller values of CWmin to high priority classes can ensure that high priority classes obtain more TXOPs than low priority ones.
- $CW_{\text{max}}[AC]$: maximal CW value for a given AC. Similar to CWmin, CWmax is also on a per AC basis.
- $AIFS[AC]$: arbitration interframe space. Each AC starts its backoff procedure after the channel is idle for a period of $AIFS[AC]$ instead of DIFS.
- $TXOP_{\text{limit}}[AC]$: the limit of consecutive transmission. During a TXOP, a station is allowed to transmit multiple data frames but limited by $TXOP_{\text{limit}}[AC]$.

![Fig. 1. The virtual backoff of the four access categories.](image-url)
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 kbps, 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 bytes for RTP, 8 bytes for UDP and 20 bytes for IP. Considering a codec packetization interval of 20 ms, the raw voice packet is 160 bytes. From the viewpoint of MAC layer, the frame payload size is 160+40=200 bytes and the data rate is 200 × 8/20 = 80 kbps.

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 link and wired links.

![A common scenario of VoIP over WLANs.](image)

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 zero-percent packet loss. The study by Garg and Kappes in [7] shows that the packet loss rate of VoIP traffic due to the overflow of the queue is very small in the cases that 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 in [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 $11\text{Mbps}/(2 \times 80\text{kbps}) = 68$, where a two-way voice communications contains two VoIP streams. However, in practice, the number of VoIP sessions that can be supported by the existing WLANs is much lesser. This is mainly due to the VoIP connection bottleneck at the AP resulting under utilization 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. Since EDCA does not differentiate between stations and the AP, the AP competes equally and shares fairly with all other stations for the channel bandwidth usage. Unaware that AP requires $N$ times more bandwidth than each VoIP local client to handle $N$ simultaneous VoIP sessions, congestion occurs at 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 $2 \times 80\text{kbps}$ 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 respectively. These analytical results are obtained from the model detailed in the next section. 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 decreases significantly. This analysis is similar to previous simulation findings [7], [8], [9] that as the number of VoIP sessions increases beyond 11, the channel 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, while the AP has become saturated, the channel has yet to reach its saturation point since the uplink transmission throughput does not drop until when there are over 20 VoIP sessions. However, beyond 11 VoIP sessions, the AP fails to capture 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 EDCA specified in the IEEE 802.11e standard. Although the standard defines service differentiation for various traffic types, there is no service differentiation among communications devices including stations and APs. Lacking service differentiation for different devices leads to the connection bottleneck at the AP discussed in the previous section.

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 EDCA parameters to the VoIP performance and identify a set of EDCA parameters that achieves improved voice capacity.

There has been a number of EDCA models [14], [15], [16], [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 non-saturation traffic description is necessary for our study. Based on the Bianchi’s original work [18], we enhance it to include EDCA operation and error prone channel consideration.

A. Markov Chain Model

Suppose there are $N$ stations and one AP. We use two ACs, AC[up] and AC[down], for the uplink traffic at the stations and the downlink traffic at the AP, respectively. For simplicity, we assume $AIFS[up] = AIFS[down]$ 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 time 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, $k$ denotes the backoff counter. Similar to the approach used in [19] for the non-saturation 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 non-saturation 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 non-saturation load adjustment of a station. Setting this probability to zero reduces the Markov Chain model to that of the saturation load condition.

![Markov Chain Model Diagram](image)

We assume that $P_{i,f}$, the unsuccessful transmission probability of AC[i], and $P_{i,b}$, 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 two parts: the collision probability $P_l$ and the failed transmission probability $P_e$ due to transmission errors. Mathematically, $P_{i,f}$ can be expressed as

$$P_{i,f} = 1 - (1 - P_l)(1 - P_e) = P_l + P_e - P_lP_e,$$

and $P_e$ is calculated by

$$P_e = 1 - (1 - \epsilon)^l,$$

where $\epsilon$ is the channel bit error rate (BER) and $l$ is the frame length in bits. Note that here we assume the collision probability $P_l$ and the frame error rate (PER) $P_e$ are independent and the bit errors are memoryless. Setting $P_e$ to zero to reduce Markov Chain model to that with perfect channel conditions.

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

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

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

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

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

and

\[ b_{i,j,k} = \frac{W_{i,j} - k}{(1 - P_{i,b})W_{i,j}} b_{i,j,0}, \quad j \in [0, m_i + h_i], \quad k \in [1, W_{i,j} - 1]. \quad (6) \]

In this way, all the values of \( b_{i,j,k} \) can be expressed in terms of \( P_{i,b}, P_{i,f} \) and \( b_{i,0,0} \). Since 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

\[ A = (1 - 2P_{i,b})(1 - 2P_{i,f})(1 - P_{i,f}^{m_i+h_i+1}) \\
+ W_i(1 - (2P_{i,f})^{m_i+1})(1 - P_{i,f}) \\
+ W_i(2P_{i,f})^{m_i}P_{i,f}(1 - P_{i,f}^{h_i+1})(1 - 2P_{i,f}) \]

From Eq. (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 \( \tau_i \) that one AC[i] tries to access the medium. It is clear that \( \tau_i \) should be equal to all the steady state probabilities of the states \( \{i,j,0\}, j = 0, 1, \ldots, m_i + h_i \), 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 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, \( P_{i,b} \), the probability that the channel is observed busy by AC[i] can be derived as

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

As for calculating \( P_{i,f} \) defined in Eq. (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, denoted as \( S_i(N) \) (\( i \in \{up, dw\} \)). Since each station and the AP operate according to the state transition diagram shown in Fig. 4, we consider the time period that all the ACs remain in their states as a time interval. In 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.,

\[ S_i(N) = \frac{R}{E[\text{time for successful transmission in an interval}]} \cdot \frac{\sum_{j=0}^{m_i+h_i} P_{i,j}E[P]}{E[I] + E[NC] + E[C]} \quad (10) \]

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 three parts: \( E[I] \), the expected value of idle time before a transmission, \( E[NC] \), transmission time without collision, and \( E[C] \), collision time.

The successful transmission probability \( P_{i,s} \) of the station and the AP can be calculated as:

\[ P_{i,s} = \begin{cases} 
\frac{\tau_{up}(1 - \tau_{up})^{N-1}(1 - \tau_{down})}{1 - P_{e}}, & \text{if } i = up \\
\frac{\tau_{down}(1 - \tau_{up})^{N}}{1 - P_{e}}, & \text{if } i = dw 
\end{cases} \quad (11) \]

where \( P_e \), the channel busy probability, is defined as

\[ P_e = 1 - (1 - \tau_{up})^{N}(1 - \tau_{down}). \quad (12) \]

Note that \( P_e \) is different from \( P_{i,b} \) in Eq. (9). \( P_{i,b} \) is the channel busy probability observed by one AC[i] while \( P_e \) is 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 \( (1 - P_{i,b}) \sigma \), where \( \sigma \) is the value of one system slot. The transmission time without collision, \( E[NC] \), includes two parts: the successful transmission time, \( E[S] \), and the failure transmission time only due to transmission errors, \( E[TE] \). We can derive \( E[S] \) and \( E[TE] \) as

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

where \( T_s \) and \( T_e \) are the average time of a successful transmission and the average time of a non-collision failure transmission, \( P_{up,s} \) and \( P_{dw,s} \) are given in Eq. (11), and the probabilities of non-collision failure transmission, \( P_{up,e} \) and \( P_{dw,e} \), are given by

\[ P_{i,e} = \begin{cases} 
\frac{\tau_{up}(1 - \tau_{up})^{N-1}(1 - \tau_{down})}{1 - P_{e}}, & \text{if } i = up \\
\frac{\tau_{down}(1 - \tau_{up})^{N+1}}{1 - P_{e}}, & \text{if } i = dw 
\end{cases} \quad (14) \]

The collision time \( E[C] \) can be expressed as

\[ E[C] = P_eT_c. \quad (15) \]

where \( T_c \) is the average collision time, \( P_e \) is the collision probability and \( P_e = 1 - NP_{up,s} - P_{dw,s} - NP_{up,e} - P_{dw,e} \).

The values of \( T_s, T_e \) and \( T_c \) in Eq. (13) and Eq. (15) depend on the channel access mode. In 802.11e standard, the channel access mode is more complicated than that of the legacy 802.11 standard. In particular, in addition to the the basic access mode and the RTS/CTS 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

\[ \begin{align*}
T_s &= AIFS + H + E[P] + SIFS + \delta + ACK + \delta \\
T_e &= AIFS + H + E[P] + \delta
\end{align*} \]

where \( ACK \) is the time duration of an ACK frame, \( H = PHY_{hdr} + MAC_{hdr} \) is the header duration, \( \delta \) is the propagation delay, and the payload length \( E[P] \) is the packet length of voice packets. As for \( T_c \), we assume transmission errors
We consider a scenario of $N$ 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 $U_{up}(N)$ and $U_{dw}(N)$ to be the throughput of (in kbps) of a particular VoIP session in the uplink and downlink transmissions respectively. Given that each VoIP session produces $2 \times 80$ kbps, then the numerical computation for $U_{dw}(N)$ requires that

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

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

$$U_{up}(N) = \begin{cases} 80, & S_{up}(N) \geq S_{up}(N) = 80 \\ S_{up}(N), & \text{otherwise} \end{cases} \quad (17)$$

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

In order 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 gradually increases its traffic load toward its saturation. In the analytical model, this is done by increasing the traffic load through adjusting the quantity $q_{up}$. In the extreme where $q_{up} = 0$, the model is reduced to the case of that saturated 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 saturated load (i.e. $q_{up} = 0$), 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.

V. SELECTING EDCA PARAMETERS

As mentioned at the beginning of Section IV, in order to overcome the bottleneck problem of VoIP, we propose to apply EDCA to provide different services to the AP and the stations. The key challenge is to select the optimal EDCA parameters. Setting $q_i = 0$, of 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 is able to tradeoff between the uplink saturation throughput and the downlink saturation throughput. Note that, in Fig. 6, when the number of stations is small, the throughput is larger than the input traffic load, which is not realistic. This is because the throughput 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 more and more close to the aggregated uplink throughput. Particularly, the cross point between the aggregated uplink throughput and the downlink throughput moves to the upper right. By continuously decreasing $W_{dw}$, we can make the cross point meets 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 be still under unsaturation conditions, we know the satisfactory of the throughput requirement can
be guaranteed. If we further move the cross point, both the downlink saturation throughput 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 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 the $W_{dw}$ from 1 to 32. It illustrates that the maximum VoIP sessions supported in the WLANs is changed under different values of $W_{dw}$ and 3 is the optimal $W_{dw}$ value. Combining with other EDCA parameters, even larger voice capacity can be achieved. Our solution can be regarded as a lower bound of the maximal voice capacity.

By considering $W_{dw} = 3$ 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 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.

We further conduct the NS-2 simulation to measure the QoS performance of 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 AP, which is calculated every 1 s. Although the throughput has a little oscillation around 80 kbps, it basically satisfy the requirement of VoIP sessions. Fig. 10 shows the delay of one VoIP session at the stations and the AP. Note that the delay includes both access delay and queueing delay. Fig. 11 plots the corresponding cumulative distribution function of the delay. We can see that most voice packets experience 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 pointed out that, for VoIP over infrastructure WLANs, the AP is the bottleneck that limits the VoIP capacity. We have proposed to use 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 non-saturation conditions. The analytical performance on saturation throughput for multi-class traffic has been validated via NS-2 simulations. We have further proposed to bypass the unsaturation performance analysis and use the developed saturation model to choose the EDCA parameters. Our analytical results have demonstrated that by only adjusting one EDCA parameter, i.e. $W_{dw}$, our proposed method improves the VoIP capacity by 20%–30%. The NS-2 simulations results have further proved that the analytically selected EDCA parameters can satisfy the VoIP QoS requirements.

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

Fig. 10. The delay of one VoIP session. Top (a): at a station. Bottom (b): at the AP.

REFERENCES


Fig. 11. The cumulative distribution function (CDF) of the VoIP delay at the stations and the AP.