Voice Capacity Analysis of WLAN With Unbalanced Traffic

Lin X. Cai, Xuemin (Sherman) Shen, *Senior Member, IEEE*, Jon W. Mark, *Life Fellow, IEEE*, Lin Cai, *Member, IEEE*, and Yang Xiao, *Senior Member, IEEE*

***Abstract*—An analytical model to study the performance of wireless local area networks (WLANs) supporting asymmetric nonpersistent traffic using the IEEE 802.11 distributed coordination function mode for medium access control (MAC) is developed. Given the parameters of the MAC protocol and voice codecs, the voice capacity of an infrastructure-based WLAN, in terms of the maximum number of voice connections that can be supported with satisfactory user-perceived quality, is obtained. In addition, voice capacity analysis reveals how the overheads from different layers, codec rate, and voice packetization interval affect voice traffic performance in WLANs, which provides an important guideline for network planning and management. The analytical results can be used for effective call admission control to guarantee the quality of voice connections. Extensive simulations have been performed to validate the analytical results.**

***Index Terms*—IEEE 802.11 DCF, unbalanced traffic, voice capacity.**

I. Introduction

VOICE over Internet protocol (VoIP) is one of the fastest growing Internet applications. It is anticipated that VoIP will be a viable alternative to the traditional public switched telephone networks (PSTNs) because of its high resource utilization and low cost. Meanwhile, the IEEE 802.11 wireless local area network (WLAN) standard is widely deployed for Internet access. Although existing WLAN applications are mainly data centric, there is a growing demand for real-time voice services over WLAN. Driven by these two popular technologies, VoIP over WLAN (VoWLAN) has been emerging as an infrastructure to provide low-cost wireless voice services.

The IEEE 802.11 standard defines two modes of medium access control (MAC) protocol, namely: 1) mandatory distributed coordination function (DCF) mode and 2) optional point coordination function (PCF) mode. Although the PCF mode is designed for real-time traffic [1], [2], it is not widely deployed due to its inefficient polling schemes, limited quality of service (QoS) provisioning, and implementation complexity.[1] On the other hand, supporting voice traffic over WLANs using the DCF mode poses significant challenges, because the performance characteristics of their physical and MAC layers are much worse than their wireline counterparts. The voice capacity of a WLAN, which is defined as the maximum number of voice connections that can be supported with satisfactory user-perceived quality, has been actively investigated both experimentally and analytically.

The voice quality and capacity of WLANs in the presence of background data traffic has been measured in [4] using a test bed consisting of commercially available components. Measurements of voice capacity with the voice codec G.711 and a 10-ms packetization interval have been carried out in [5]. Because experimental results cannot predict voice capacity when new wireless technologies or voice codecs are emerged and experiments alone cannot fully reveal the relationship between voice capacity and system parameters, the voice capacity of a WLAN has been theoretically estimated in [6] based on the assumptions that there is no collision during transmissions and all mobile stations take advantage of the backoff time of the access point (AP) to fulfill their own backoff requirements. An analytical model to estimate the voice capacity of the IEEE 802.11a/b-based WLANs is developed in [7]. It is assumed that there are always two and only two active stations competing for the wireless channel, where an active station is defined as a station that currently has a frame in service. The voice capacity obtained in [6] and [7] may be overly optimistic due to these simplified assumptions. A loose estimation of voice capacity is harmful for admission control, because once traffic load exceeds the network capacity, the quality of all ongoing voice traffic will be jeopardized.

In addition to voice capacity analysis, the throughput and delay of traffic over a WLAN have been extensively studied in the literature. Bianchi [8] develops a bidimensional discrete-time Markov chain model to calculate the system throughput as a function of the number of saturated stations. Here, a saturated station always has a frame ready for transmission. In reality, some stations, especially those with real-time voice traffic, are unsaturated. When there are more than one station sharing the wireless resource, the maximum throughput of WLAN

Manuscript received September 17, 2005; revised November 28, 2005 and December 2, 2005. This work was supported by the Natural Sciences and Engineering Research Council of Canada under Strategic Grant STPGP 257682 and by a research grant from Bell University Laboratories under the sponsorship of Bell Canada. The review of this paper was coordinated by Prof. Y. Fang.

L. X. Cai, X. Shen, and J. W. Mark are with the Centre for Wireless Communications, Department of Electrical and Computer Engineering, Faculty of Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: lcai@bbcr.uwaterloo.ca; xshen@bbcr.uwaterloo.ca; jwmark@bbcr.uwaterloo.ca).

L. Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P5C2, Canada (e-mail: cai@uvic.ca).

Y. Xiao is with the Department of Computer Science, University of Memphis, Memphis, TN 38152-3240 USA (e-mail: yangxiao@ieee.org).

Digital Object Identifier 10.1109/TVT.2006.874145

[1] Because both DCF and PCF have limited support for real-time applications, the IEEE 802.11e has been proposed to enhance the current 802.11 MAC to support applications with stringent QoS requirements [3], but it is unclear if and when the standard will be widely deployed.

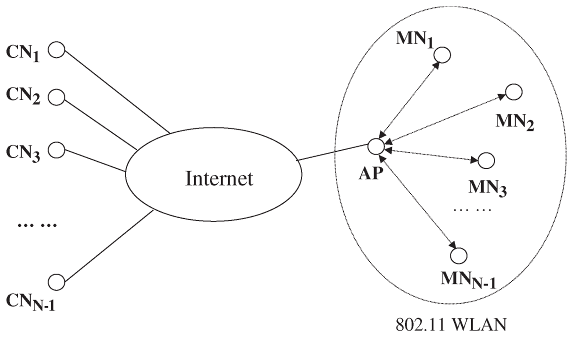

Fig. 1. Network scenario.

can be achieved only in the unsaturated station case [18]. Therefore, Bianchi's model may not be directly applicable to voice capacity analysis. Tickoo and Sikdar [9], [10] model the queue of an unsaturated station as a discrete-time G/G/1 queue and use this model to analyze the frame delay distribution. Both Bianchi's and Tickoo's models consider a homogeneous scenario in an independent (or *ad hoc*) basic service set (BSS), and they assume that all stations have the same traffic load and frame service rate. However, the majority of existing WLANs are set up in infrastructure mode, where mobile stations access the Internet through an AP, which coordinates all traffic to and from the WLAN. In an infrastructure-based WLAN, the AP has a much higher traffic load and is the bottleneck. The unbalanced traffic load affects network performance and voice capacity, which needs further investigation.

In this paper, we study the network performance of an infrastructure-based WLAN, considering the practical issue induced by unbalanced traffic. The network scenario is shown in Fig. 1, where the WLAN is composed of one AP and $N-1$ mobile nodes (MNs). The AP and $N-1$ correspondent nodes (CNs) are connected to a wired backbone network. Voice connections are established between MNs and CNs through the AP. By considering the different traffic loads of the stations, we obtain the conditional collision probabilities and the frame service rates of the AP and MNs, respectively. We further analyze the queue utilization ratio (or traffic intensity) of the AP and MNs. Given a voice codec, an accurate upper bound on the number of simultaneous voice connections that can be supported in an infrastructure-based WLAN is obtained.

Our main contribution has two aspects, namely: 1) development of an analytical model for studying the system performance of WLANs with asymmetric traffic and unsaturated stations and 2) use of the model to quantify the voice capacity of infrastructure-based WLANs. The analytical results reveal how unbalanced traffic, overheads from different layers, codec rate, and voice packetization interval affect the voice capacity. This information provides a useful guideline for radio resource allocation and management and offers insights for future network planning and protocol design. The analytical results are validated by simulations using the NS-2 (version 2.27) [11]. Although this paper focuses on the basic access mode of DCF, the approach can be extended to the request-to-sender/clear-to-send (RTS/CTS) mode and is suitable for analyzing the voice capacity of any wireless networks using the carrier sense multiple access/collision avoidance (CSMA/CA) MAC protocol.

The remainder of this paper is organized as follows: The legacy IEEE 802.11 MAC protocol and VoIP system are presented in Section II. The analytical model, which is developed in Section III, is used to derive the voice capacity of WLANs in Section IV. Section V presents the simulation results, followed by concluding remarks in Section VI.

II. Background

A. IEEE 802.11 DCF-Based MAC

The IEEE 802.11 DCF-based MAC protocol uses the CSMA/CA mechanism [12]. A station monitors the medium before attempting transmission. If the medium is sensed busy, the station defers transmission until the medium is sensed idle for a period of time equal to a DCF interframe space (DIFS). After the DIFS medium idle time, it enters the backoff phase in which it sets a random backoff counter randomly chosen from $[0, \mathrm{CW})$, where CW is the contention window size. The backoff counter decreases by one for every time slot if the medium is idle; otherwise, the counter freezes, and the decrement resumes after the medium is sensed idle again for a DIFS. When the backoff counter reaches zero, the station transmits the frame. If another station transmits a frame at the same time, a collision occurs, and both transmissions fail. CW is doubled after a collision until it reaches the maximum value $(\mathrm{CW}_{\max})$, and the sender reschedules the transmission by randomly choosing a backoff counter in $[0, \mathrm{CW})$. The frame is dropped when the retransmission limit is reached. After a successful transmission, CW is reset to its minimum value $(\mathrm{CW}_{\min})$. Upon receiving a frame successfully, the receiver transmits an acknowledgment (ACK) following a short interframe space (SIFS). Two medium access techniques are specified in DCF, namely 1) basic access mechanism and 2) RTS/CTS mechanism. Frames are transmitted using the RTS/CTS mechanism if their payload exceeds a given threshold; otherwise, the basic access is used. Voice frames are transmitted using the basic access mechanism because of their small payload size. Furthermore, without link layer fragmentation, one voice packet corresponds to one link layer frame.

With the IEEE 802.11 DCF-based MAC, all have the same priority to access the channel. This is unfavorable to the AP, which has a much higher traffic load. In addition, the CSMA/CA mechanism was originally designed for data transmission, without considering delay-sensitive voice traffic. Before being successfully transmitted, each frame has to wait a random time period, which depends on the network load and collisions that it experienced. A high collision probability reduces the frame service rate and accentuates the queue length and delay, which should be avoided for voice traffic. Thus, it is critical to obtain the upper bound of traffic load to limit the contention and collisions, which will be discussed in Section IV.

B. VoIP System

VoIP has been widely accepted for its cost effectiveness and easy implementation. A VoIP system consists of three

TABLE I
FREQUENTLY USED VOICE CODECS

Voice Codec		G.711	G.723a	G.729	iLBC
Codec Bit Rate		(64kbps)	(5.3/6.3kbps)	(8kbps)	(15.2/13.3kbps)
Sample Period	Arrival Rate (frames/sec)	Payload (Byte)	Payload (Byte)	Payload (Byte)	Payload (Byte)
10ms	100	80		10	
20ms	50	160		20	38
30ms	33.33	240	20/24	30	50
40ms	25	320		40	
50ms	20	400		50	
60ms	16.67	480	40/48	60	

indispensable components, namely 1) codec, 2) packetizer, and 3) playout buffer. Analog voice signals are digitized, compressed, and encoded into digital voice streams by the codecs. The output digital voice streams are then packed into constant-bit-rate (CBR) voice packets by the packetizer. Each voice packet has a 40-B real-time transport protocol (RTP)/user datagram protocol (UDP)/IP header. After voice packets are delivered through the network, the reverse process of decoding and depacketizing is accomplished at the receiver. A two-way conversation is very sensitive to packet delay jitter, but it can tolerate a certain degree of packet losses. Therefore, a playout buffer is used to smooth the speech by eliminating the delay jitter at the receiver. Any packets arriving later than the playout time will simply be discarded. Other components, such as voice activity detector (VAD), loss/error concealment, and echo control, etc., are also included in the system to enhance the functionality and performance of a VoIP system [13], [14].

Table I lists the main attributes of some frequently used voice codecs with different packetization intervals. Different codecs use different compression algorithms, resulting in different bit rates. G.711 is the international standard for encoding telephone audio, which has a fixed bit rate of 64 kb/s. If the packetization interval is 20 ms, which corresponds to a rate of 50 packets/s, the payload size will be $64\,000/(50 * 8) = 160$ B. If the packetization interval is reduced to 10 ms, which corresponds to a rate of 100 packets/s, the payload size will be reduced to $64\,000/(100 * 8) = 80$ B. G.723, G.729, and internet low bit-rate codec (iLBC) are popular codecs used by VoIP applications. They have lower bit rates at the expense of higher codec complexity. G.723 is one of the most efficient codecs with the highest compression ratio and is usually used in video conferencing applications. G.729 is an industry standard with high bandwidth utilization for toll-quality voice calls. iLBC is developed for robust voice communications that can achieve a graceful degradation of voice quality with severe packet losses [15]; it has been chosen by many Internet softphone applications, e.g., Skype. iLBC has a codec bit rate of 13.3 kb/s for a 30-ms packetization interval and 15.2 kb/s for a 20-ms interval.

III. Analytical Model

In this section, we present an analytical model for studying the performance of WLANs with asymmetric traffic using the DCF mode for MAC. The analytical model is used to derive the voice capacity in Section IV.

We consider a single-hop fully connected WLAN with N stations, and every mobile station can sense the status of the shared wireless channel. Time is discretized into slots, and all stations are synchronized to operate in slotted time. The wireless channel is assumed ideal such that all transmitted frames can be received error free if there is no collision. Define the conditional collision probability p_i as the probability of a collision seen by a frame being transmitted by the tagged station i. p_i is assumed constant and independent of the number of retransmissions the packet has experienced. Because the probability of three or more stations simultaneously transmitting is very small, in what follows, we assume that collisions are due to two stations transmitting simultaneously.

Let the traffic arrival rate and the frame service rate of station i be denoted as λ_i and μ_i frames per slot, respectively, where $i = 0, 1, \ldots, N-1$. The queue utilization ratio of station i is $\rho_i = \lambda_i/\mu_i$. All frames are transmitted at the same transmission rate, and our analysis can be extended to consider rate adaptation schemes.

Define $p_i[T]$ as the probability that station i transmits a frame in a randomly chosen slot. Conditional on the queue state, the transmission probability of station i can be derived as

$$p_i[T] = p_i[T|\mathrm{QE}]p_i[\mathrm{QE}] + p_i[T|\mathrm{QNE}]p_i[\mathrm{QNE}] \tag{1}$$

where $p_i[\mathrm{QE}]$ and $p_i[\mathrm{QNE}]$ are the probabilities of an empty queue and a nonempty queue of station i, respectively.

The queue of a station is considered empty when the station is idle, i.e., no frame is in service or waiting for service. A station is idle with a probability of $1 - \rho_i$ for $\rho < 1$. Therefore, $p_i[\mathrm{QE}] = 1 - \rho_i$ and $p_i[\mathrm{QNE}] = \rho_i$. Because a station never transmits with an empty queue, $p_i[T|\mathrm{QE}] = 0$. Defining $\tau_i = p_i[T|\mathrm{QNE}]$ to simplify the notation, the transmission probability of station i is given by

$$p_i[T] = 0 * (1 - \rho_i) + \tau_i * \rho_i = \rho_i \tau_i. \tag{2}$$

Here, we only consider the unsaturated case where $\rho_i < 1$.

If station i transmits in a given slot, a collision occurs if at least one of the remaining stations also transmits in the same slot. We have

$$p_i = 1 - \prod_{j=0, j\neq i}^{N-1} (1 - p_j[T]) = 1 - \prod_{j=0, j\neq i}^{N-1} (1 - \lambda_j \tau_j / \mu_j) \tag{3}$$

where $i = 0, 1, \ldots, N-1$.

Conditional on a nonempty queue, the transmission probability of station i can be approximated as

$$\tau_i = \frac{E[M_i]}{\overline{w_i}} \tag{4}$$

where $\overline{w_i}$ is the average backoff time for station i to successfully transmit a packet, and $E[M_i]$ is the average number of transmission attempts station i made during $\overline{w_i}$. Each transmission attempt has a collision probability of p_i and a success probability of $1 - p_i$. With the IEEE 802.11 DCF mode, a backoff counter is uniformly chosen over $[0, \mathrm{CW}]$, where CW is the current contention window size. The exponential backoff procedure of a station can be modeled as a geometrically distributed random variable. Thus, the average backoff time of station i can be derived as

$$\begin{aligned}\overline{w_i} = (1-p_i)\frac{W}{2} + \cdots + p_i^{m'}(1-p_i)\frac{\sum_{i=0}^{m'} 2^i W}{2} + \cdots \\ + p_i^m \frac{\sum_{i=0}^{m'} 2^i W + (m-m')2^{m'} W}{2}\end{aligned} \tag{5}$$

where m' is the maximum backoff stage, m is the retransmission limit, and W is the minimum backoff window size. According to the IEEE 802.11 standard, m' is 5, m is 7, and W is 32 time slots. Similarly, the transmission attempts of station i can also be modeled as a geometrically distributed random variable, and the average number of transmission attempts of station i can be derived as

$$\begin{aligned}E[M_i] &= (1-p_i)\cdot 1 + \cdots + p_i^m \cdot (m+1) \\ &= \frac{1-p_i^{m+1}}{1-p_i}.\end{aligned} \tag{6}$$

Given the system parameters defined by the standard, $E[M_i]$ and $\overline{w_i}$ are determined by p_i alone. Therefore, by substituting (5) and (6) into (4), τ_i can be represented as a function of p_i.

To determine p_i and ρ_i, we need to obtain the average service time of a frame $1/\mu_i$, which is the time interval between the time instant that a frame is ready to be transmitted by station i and the time instant that the frame is successfully transmitted. During $1/\mu_i$, in addition to a successful transmission by the tagged station i, the following events may occur:

1) successful transmissions by the remaining $N-1$ stations;
2) collisions;
3) channel idleness when station i is in its backoff stage(s).

The transmission time of the frame being sent by i is T_{s_i}, which is the time duration the channel is sensed busy due to a successful transmission by station i. We study the system when it is operating in stable state, i.e., when all incoming packets are transmitted within a finite delay. During $1/\mu_i$, on the average, the remaining stations successfully transmit $(1/\mu_i)\sum_{j=0,j\neq i}^{N-1} \lambda_j$ frames, which contribute to $(1/\mu_i) \times \sum_{j=0,j\neq i}^{N-1} \lambda_j T_{s_j}$ time slots. Before the stations successfully transmit the frames, the total amount of collision time that each station experiences is $(1/\mu_i)\sum_{j=0,j\neq i}^{N-1} \lambda_j \overline{T_{c_j}} + \overline{T_{c_i}}$, where $\overline{T_{c_i}}$ is the average collision time of a frame transmitted by station i. Denote T_{c_i} as the collision time that station i experiences each time a collision occurs; therefore, $\overline{T_{c_i}}$ can be derived as a function of p_i, which is expressed as follows:

$$\begin{aligned}\overline{T_{c_i}} &= p_i(1-p_i)\cdot T_{c_i} + \cdots + p_i^m(1-p_i)\cdot mT_{c_i} \\ &= \frac{p_i\left(1-(m+1)p_i^m + mp_i^{m+1}\right)}{1-p_i} T_{c_i}.\end{aligned} \tag{7}$$

T_{s_i} and T_{c_i} can be obtained given the frame length of station i. Because a collision is assumed to occur due to simultaneous transmissions by two stations, the duration for the channel to be busy due to collision equals half of the total amount of collision time experienced by all stations, which is $(1/2)((1/\mu_i)\sum_{j=0,j\neq i}^{N-1} \lambda_j \overline{T_{c_j}} + \overline{T_{c_i}})$. Finally, station i spends $\overline{w_i}$ in the backoff stage before it successfully transmits the current frame. Therefore, we have

$$\begin{aligned}\frac{1}{\mu_i} = T_{s_i} + \frac{1}{\mu_i}\sum_{j=0,j\neq i}^{N-1} \lambda_j T_{s_j} \\ + \frac{1}{2}\left(\frac{1}{\mu_i}\sum_{j=0,j\neq i}^{N-1} \lambda_j \overline{T_{c_j}} + \overline{T_{c_i}}\right) + \overline{w_i}\end{aligned} \tag{8}$$

where $i = 0, 1, \ldots, N-1$.

Given the arrival rates $\vec{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_{N-1}]$, (3) and (8) can be solved numerically to obtain $\vec{p} = [p_0, p_1, \ldots, p_{N-1}]$, $\vec{\mu} = [\mu_0, \mu_1, \ldots, \mu_{N-1}]$, and $\vec{\rho} = [\rho_0, \rho_1, \ldots, \rho_{N-1}]$.

IV. Voice Capacity Analysis

As shown in Fig. 1, the WLAN consists of one AP and $N-1$ MNs. Each MN communicates with a CN via the AP. The voice stream of station i is modeled as a CBR traffic (without the use of silence suppression) with an arrival rate of λ_i frames per slot. We assume that all MNs in the WLAN use the same voice codec so they have the same traffic load and frame service rate $\lambda_i = \lambda_1, i = 1, \ldots, N-1$ and $\mu_i = \mu_1$, $i = 1, \ldots, N-1$, respectively. The CBR traffic model is used to derive the voice capacity because of two reasons, namely 1) some voice codecs do not use the silence suppression scheme and 2) if the silence suppression scheme is used and the traffic exhibits on–off characteristics, the upper bound derived using the CBR traffic model is robust in the worst case scenario when all voice flows in the “ON” state. Because the number of flows in a WLAN is relatively small, a tight upper bound considering the worst case scenario is desired.

In the infrastructure-based WLAN, all traffic to the MNs is transmitted by the AP, i.e., the traffic load of the AP is $N-1$ times that of an MN. Therefore, the traffic arrival rate of the AP is $\lambda_0 = (N-1)\lambda_1$ frames per slot. The frame service rate of the AP is denoted as μ_0 frames per slot. The queue utilization ratios at the AP and MNs are denoted by $\rho_0 = (\lambda_0/\mu_0) = (N-1)\lambda_1/\mu_0$ and $\rho_i = \lambda_1/\mu_1, i = 1, \ldots, N-1$, respectively.

According to (3), the conditional collision probability for frames being transmitted by the AP (p_0) and that for frames being transmitted by an MN (p_1) are given by

$$\begin{cases} p_0 = 1 - (1 - \rho_1 \tau_1)^{N-1} \\ p_1 = 1 - (1 - \rho_1 \tau_1)^{N-2} (1 - \rho_0 \tau_0) \end{cases}. \tag{9}$$

From (4)–(6), τ_0 and τ_1 are functions of p_0 and p_1, respectively, which yields

$$\begin{cases} \tau_0 = E[M_0] / \overline{w_0} \\ \tau_1 = E[M_1] / \overline{w_1} \end{cases} \tag{10}$$

where

$$\begin{cases} E[M_0] = \dfrac{1 - p_0^{m+1}}{1 - p_0} \\ E[M_1] = \dfrac{1 - p_1^{m+1}}{1 - p_1} \\ \overline{w_0} = (1 - p_0) \dfrac{W}{2} + \cdots + p_0^{m'} (1 - p_0) \dfrac{\sum_{i=0}^{m'} 2^i W}{2} \\ \qquad + \cdots + p_0^{m} \dfrac{\sum_{i=0}^{m'} 2^i W + (m - m') 2^{m'} W}{2} \\ \overline{w_1} = (1 - p_1) \dfrac{W}{2} + \cdots + p_1^{m'} (1 - p_1) \dfrac{\sum_{i=0}^{m'} 2^i W}{2} \\ \qquad + \cdots + p_1^{m} \dfrac{\sum_{i=0}^{m'} 2^i W + (m - m') 2^{m'} W}{2}. \end{cases}$$

Because all stations use the same voice codec, all of the voice frames have the same size. Denote T_s as the time duration when the channel is sensed busy because of a successful transmission and T_c as the time duration when the channel is sensed busy due to failed transmissions. In the basic access mode, T_s consists of the transmission time for the voice frame, including the headers encapsulated in each layer, an SIFS, the transmission time of an ACK frame, and a DIFS and is expressed as

$$T_s = T_{\mathrm{data}} + \mathrm{SIFS} + T_{\mathrm{ACK}} + \mathrm{DIFS}. \tag{11}$$

T_c consists of the transmission time for a voice frame, the time waiting for an ACK timeout, and a DIFS and is expressed as

$$T_c = T_{\mathrm{data}} + \mathrm{ACK}_{\mathrm{timeout}} + \mathrm{DIFS}. \tag{12}$$

The average collision time of a frame transmitted by the AP and by an MN can be derived from (7) as

$$\begin{cases} \overline{T_{c_0}} = \frac{p_0 \left[1 - (m+1) p_0^{m} + m p_0^{m+1}\right] T_c}{1 - p_0} \\ \overline{T_{c_1}} = \frac{p_1 \left[1 - (m+1) p_1^{m} + m p_1^{m+1}\right] T_c}{1 - p_1}. \end{cases} \tag{13}$$

From the time an AP transmits a frame until the frame is transmitted successfully, the time interval $1/\mu_0$ consists of four parts, namely 1) on the average, the remaining $N - 1$ MNs successfully transmit $(N - 1)\lambda_1/\mu_0$ frames, which contribute $(N - 1)\lambda_1 T_s/\mu_0$, 2) AP spends T_s in transmitting the current frame, 3) before the stations successfully transmit these frames, the total time that the channel is sensed busy due to failed transmissions is $[(N - 1)\lambda_1 \overline{T_{c_1}}/(2\mu_0) + \overline{T_{c_0}}/2]$, and 4) $\overline{w_0}$ is the average backoff time the AP experiences before it successfully transmits the current frame.

Similarly, the time interval $1/\mu_1$ also consists of four parts, namely 1) remaining $N - 2$ MNs and the AP contribute $(N - 2)(\lambda_1/\mu_1)T_s$ and $((N - 1)\lambda_1/\mu_1)T_s$ in successful transmissions, respectively, 2) tagged MN spends T_s in transmitting the current frame, 3) collision time is $(1/2)[(N - 2)(\lambda_1/\mu_1)\overline{T_{c_1}} + \overline{T_{c_1}} + ((N - 1)\lambda_1/\mu_1)\overline{T_{c_0}}]$, and 4) average backoff time of the tagged station is $\overline{w_1}$. Therefore, the average service time for the AP and the MNs are given as follows:

$$\begin{cases} \dfrac{1}{\mu_0} = \left((N - 1)\dfrac{\lambda_1}{\mu_0} + 1\right) T_s + \overline{w_0} \\ \qquad + \dfrac{1}{2}\left((N - 1)\dfrac{\lambda_1}{\mu_0}\overline{Tc_1} + \overline{Tc_0}\right) \\ \dfrac{1}{\mu_1} = \left((N - 2)\dfrac{\lambda_1}{\mu_1} + 1 + \dfrac{(N - 1)\lambda_1}{\mu_1}\right) T_s + \overline{w_1} \\ \qquad + \dfrac{1}{2}\left(\left((N - 2)\dfrac{\lambda_1}{\mu_1} + 1\right)\overline{Tc_1} + \dfrac{(N - 1)\lambda_1}{\mu_1}\overline{Tc_0}\right). \end{cases} \tag{14}$$

Note that we only consider the frames that have been received successfully. The service time of frames being dropped after m failed retransmissions is not included in the aforementioned equations. In general, the frame drop probability $p_{\mathrm{drop}} = p_i^{m+1}$ is negligible when p_i is small and m is large. Equations (9) and (14), along with (10)–(13), can be solved numerically to obtain p_0, p_1, μ_0, μ_1, ρ_0, and ρ_1.

A station is considered stable only if its queue utilization ratio $\rho_i < 1$, i.e., the traffic arrival rate is less than the frame service rate. In an infrastructure-based WLAN, the AP is the bottleneck because the traffic to all MNs has to go through the AP. Therefore, the maximum number of voice connections that can be accommodated in a WLAN can be obtained under the constraint that the AP is stable, i.e., the queue utilization ratio of the AP $\rho_0 < 1$. In addition, the number of active stations in the WLAN can also be obtained as $\sum_{i=0}^{N-1} \rho_i = \rho_0 + (N - 1)\rho_1$.

We investigate the maximum number of VoIP connections that can be supported in a single-AP WLAN. The main parameters of the IEEE 802.11a/b and the upper layer–header overheads of the voice frames are listed in Table II. Both data and ACK frames are transmitted at the highest rate.

The 802.11b standard defines the highest rate to be 11 Mb/s. The values of a slot duration, DIFS, and SIFS are 20, 50, and 10 μs, respectively. Each ACK frame has 14 B, and it takes $14 * 8/11 = 10.2$ μs for transmission. Each data frame has a 34-B MAC layer overhead and a 40-B RTP/UDP/IP header overhead, which take $34 * 8/11 = 24.7$ μs and $40 * 8/11 = 29.1$ μs, respectively, to transmit. In addition, it takes 192 μs to transmit the physical layer overheads consisting of 48 μs

TABLE II
PARAMETERS OF VOICE OVER 802.11

		802.11b	802.11a
Highest Channel Rate		11Mbps	54Mbps
Slot Time		20 µs	9 µs
SIFS		10 µs	16 µs
DIFS		50 µs	34 µs
CW_{min}		32	16
CW_{max}		1024	1024
Retry Limit		7	7
Tvoice	PLCP & Preamble	192 µs	24 µs
	MAC Header + FCS	24.7 µs	5 µs
	RTP/UDP/IP Header	29.1 µs	6 µs
	Voice Pay load	(payload *8/11) µs	(pay load *8/54) µs
TACK	PLCP & Preamble	192 µs	24 µs
	ACK Frame	10.2 µs	2.1 µs

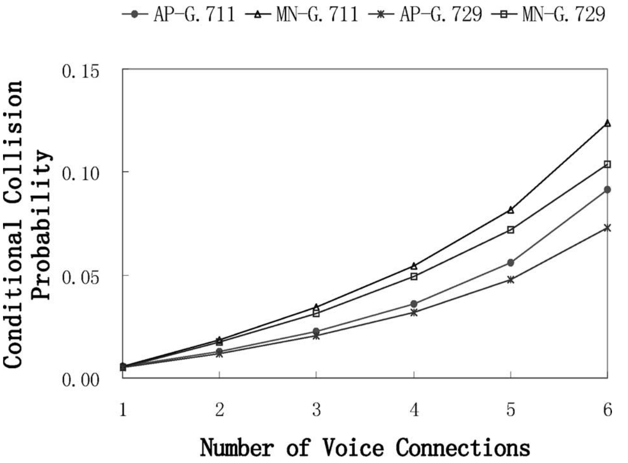

Fig. 2. Comparison of the conditional collision probabilities of the AP and MNs (802.11b).

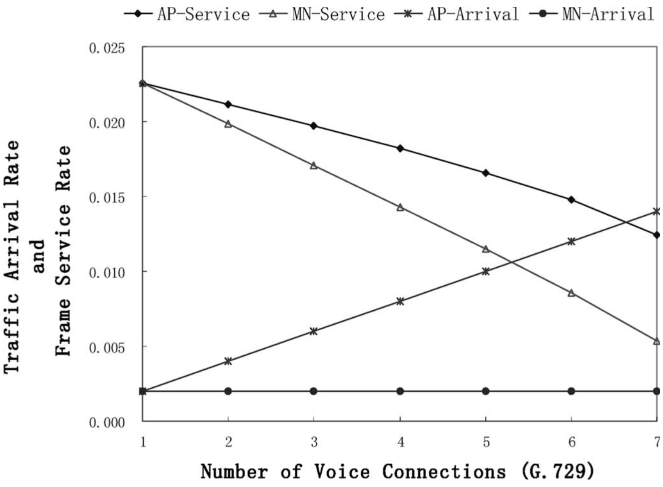

Fig. 3. Traffic arrival rate and frame service rate (802.11b).

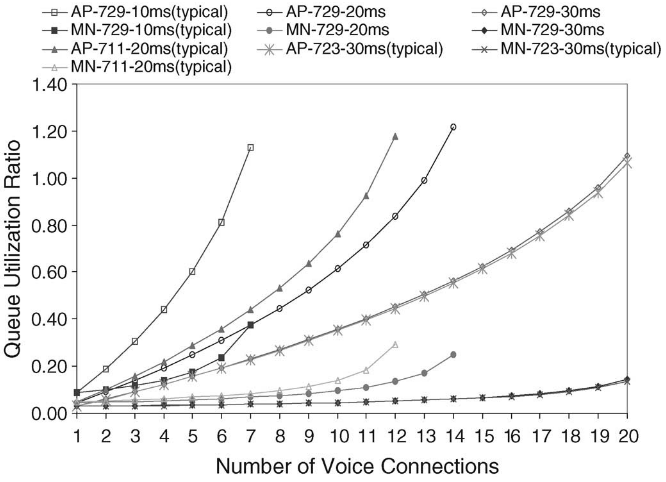

Fig. 4. Queue utilization ratio of the AP and MNs (802.11b).

physical layer convergence protocol (PLCP) header and 144 μs preamble. In the 802.11a standard, the maximum data rate is 54 Mb/s, approximately five times that of 802.11b. It takes 2.1, 5, and 6 μs to transmit the ACK frame, MAC layer overhead, and the RTP/UDP/IP headers, respectively. The values of a slot time, DIFS, and SIFS are 9, 34, and 16 μs, respectively; and it takes 24 μs to transmit the physical layer overhead, which is eight times smaller than that in 802.11b.

We use Maple 9.5 [16] to calculate the analytical results. Fig. 2 shows the conditional collision probabilities of the AP and MNs with G.711 and G.729 codecs and a 10-ms packetization interval in an IEEE 802.11b WLAN. The collisions increase with the number of voice connections. Due to the larger payload, the collision probability of G.711 is higher than that of G.729. Because the traffic load of the AP is $N-1$ times the load of an MN, collisions are more likely to occur from the viewpoint of an MN than that from the AP.

Real-time applications are very sensitive to delay and jitter. With a constant arrival rate, delay guarantee of real-time applications is possible only when the traffic arrival rate is less than the service rate $(\rho_i < 1)$. A station is considered unstable if its queue utilization ratio $\rho_i \geq 1$. For an unstable station, the queue will build up in the station, and thus, the real-time applications

TABLE III
COMPARISON OF THE MAXIMUM NUMBER OF VOIP CONNECTIONS (802.11b)

	G.711			G.729			G.723			iLBC
Audio (ms)	Proposed Analysis	[7]	[6]	Proposed Analysis	[7]	[6]	Proposed Analysis	[7]	[6]	Proposed Analysis
10	6	6	6	6	7	7				
20	11	12	12	13	14	14				12
30	15	17	18	19	21	22	19	21	22	18
40	19	21	22	25	28	28				
50	22	25	26	31	34	35				
60	25	28	29	37	41	42	37	42	42	

will be damaged due to the ever-increasing queuing delay and packet losses due to buffer overflow.

We use G.729 with a 10-ms packetization interval for illustration. Due to the characteristics of the voice traffic, the traffic arrival rate of an MN is constant. With the increase of the number of MNs, the traffic arrival rate of the AP increases linearly, whereas the frame service rate exhibits a nonlinear decreasing trend, as shown in Fig. 3. Although the frame service rate of an MN degrades more rapidly than that of the AP due to the higher collision probability, the AP enters the unstable state before the MNs because of its much higher traffic load. As shown in Fig. 3, when the seventh G.729 voice connection joins in, the queue of the AP is no longer stable. Therefore, with G.729 and a 10-ms packetization interval, at most six bidirectional VoIP connections can be supported in an IEEE 802.11b WLAN. One more VoIP connection will jeopardize the performance of all voice connections. Therefore, an accurate upper bound is critical for VoIP admission control in order to maintain an acceptable QoS for all VoIP connections.

As shown in Fig. 4, the queue utilization ratio of the AP (ρ_0) is always much higher than that of an MN (ρ_1) due to the higher traffic load. The maximum number of voice connections with $\rho_0 < 1$ can also be observed. With the G.729 codec, 6 voice connections with a 10-ms packetization interval, 13 connections with a 20-ms interval, and 19 with a 30-ms interval can be supported in an 802.11b WLAN. It can be seen that using G.729 or G.723 makes little difference on the maximum number of voice connections being supported in the WLAN. With G.729 or G.723, up to 19 simultaneous voice connections with a 30-ms packetization interval can be supported. The payload of G.711 is eight times that of G.729, but only two fewer connections can be accommodated. Compared to the huge overheads specified in the physical and MAC layers, the payload difference between different codecs is relatively small. The maximum number of connections with iLBC is similar to that with G.723 and G.729. For VoWLAN, G.729 and iLBC are preferred over G.723 because less compression is required.

Another observation is that more VoIP connections can be accommodated when the packetization interval is enlarged. However, larger packetization interval will result in longer delay. There is a tradeoff between the delay constraint and the voice capacity. In addition, when we use the short preamble of 72 b instead of the long preamble of 144 b, two more G.711 voice connections can be admitted into the network, which indicates the significant effect of the physical layer overhead.

Table III tabulates the maximum number of VoIP connections for different codecs in an 802.11b WLAN. It shows that only a very limited number of voice connections can be supported in a WLAN, even with the bandwidth efficient codec G.723. Compared to the results in [6] and [7], the obtained analytical upper bounds are much tighter. When the packetization interval is enlarged to accommodate more voice connections, the analytical results given in [6] and [7] become too optimistic. This is because in [6], it is assumed that any transmitted frame is received successfully without any collision. This assumption may not hold, especially when the number of voice connections is close to the capacity. A simple approximation that there are always two active stations (one is the AP and the other is an MN) in the network and the collision probability keeps as low as 0.03, independent of the number of voice connections, is made in [7]. However, in the unsaturated station scenario, the number of active stations is not a constant but increases with the number and the traffic intensity of the stations in the network. The AP has a frame in service with probability ρ_0, whereas each MN has a frame in service with probability ρ_1. On the average, there are $\rho_0 + (N-1)\rho_1$ stations that have a frame in service. As shown in Fig. 5, the average number of active stations in the WLAN varies from 0.02, when there is only one voice connection, to above 3, when the AP is nearly saturated.

The data rate of an 802.11a WLAN is roughly five times that of an 802.11b WLAN. However, the voice capacity of 802.11a is less than five times that of 802.11b due to the different parameter values specified in the standard, such as the minimum contention window, duration of a slot, DIFS, and SIFS. For example, a smaller minimum contention window may result in more collisions, and a larger SIFS causes longer service time. The above two parameters may reduce the voice capacity. On the other hand, a smaller physical layer overhead and shorter slot duration result in higher voice capacity. The effect of different codecs and packetization intervals on the voice capacity of an 802.11a WLAN is given in Table IV.

V. SIMULATION RESULTS

We further validate the analytical results by extensive simulations using a network simulator (NS2-2.27) [11]. We use the same parameter values of the IEEE 802.11b as those listed in Table II. The 802.11 code in NS2 is rigorously checked, and some modifications are made according to the standard: The ACK transmission rate is set to 11 Mb/s, and the preamble transmission rate is kept at 1 Mb/s. The network topology is shown in Fig. 1. In the wired network, the links connecting the AP and the CNs have a data rate of 100 Mb/s with a 20-ms propagation delay.

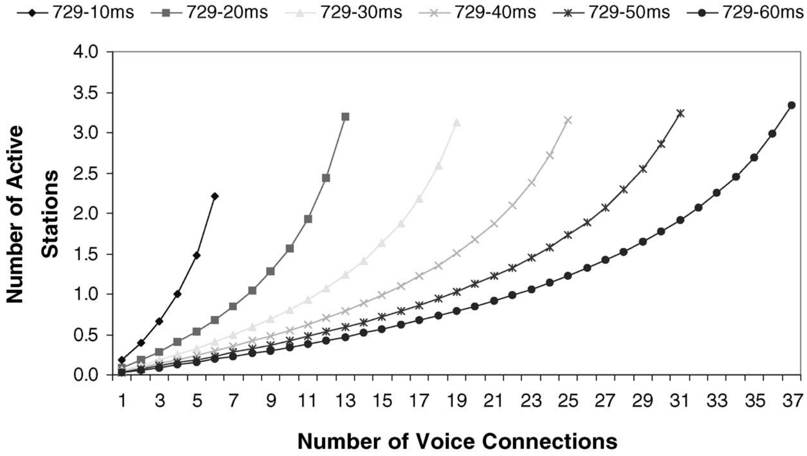

Fig. 5. Number of active stations.

TABLE IV
COMPARISON OF THE MAXIMUM NUMBER OF VOIP CONNECTIONS (802.11a)

	G.711		G.729		G.723		iLBC
Audio (ms)	Proposed Analysis	[7]	Proposed Analysis	[7]	Proposed Analysis	[7]	Proposed Analysis
10	25	30	27	32			
20	47	56	53	64			53
30	66	79	79	95	80	96	78
40	82	98	105	126			
50	97	116	130	156			
60	110	131	155	185	158	187	

The end-to-end packet delay bound is set to 150 ms to maintain good voice quality.[2] Any packets arriving after 150 ms will be discarded from the receiver's playout buffer. In order to show the queue accumulating effect in the AP, the buffer size of the AP is set to 300 packets. Initially, a voice connection is established every 10 ms to gradually approach the network capacity, with the starting time randomly chosen over [0, 10] ms. To eliminate the warming-up effects, the simulation data are collected from 10 to 100 s.

For G.729 with a 10-ms packetization interval, the packet delay of the uplink (from an MN to the AP) and downlink (from the AP to an MN) voice flow is very low when there are fewer than six connections in the WLAN. When the seventh station joins the system, the delay of the downlink flow increases rapidly, whereas the delay of the uplink is as low as 2 ms. It implies that the AP is saturated when the queue utilization ratio $\rho_0 \geq 1$. Meanwhile, the queue utilization ratio of the MNs ρ_1 is much less than 1. When more stations join, which results in more collisions in the network and decreases the frame service rate, the delay of the uplink flow also increases to more than 300 ms, implying that the MNs become saturated when there are more than 12 voice connections, as shown in Fig. 6.

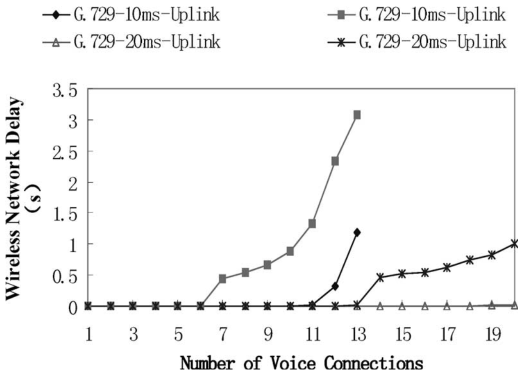

Fig. 6. Delay comparison between uplink and downlink voice flows.

Because the downlink transmissions always suffer longer queuing delay at the AP than the uplink transmissions at the MNs, we are more interested in the delay of downlink flows due to this bottleneck effect. Fig. 7 shows the delay outage ratio (the ratio of the packets with end-to-end delay exceeding 150 ms over the packets being transmitted) of downlink flows with G.729 and different packetization intervals. Due to the nonbursty characteristics of voice traffic, packet delay is quite low, and no packet is discarded from the playout buffer when all stations are not saturated. However, the outage ratio of downlink flows becomes significant when the AP is saturated due to the ever-increasing queuing delay at the AP.

In the simulation, the maximum number of voice connections is obtained in the way that one more connection will result in the delay outage ratio larger than 1%. As shown in Fig. 8, the simulation results conform with our analysis results quite well, and the obtained upper bounds are more accurate than the result in [6] and [7], because we consider the different collision probabilities and queue states of the AP and MNs. Therefore,

[2] The International Telecommunication Union (ITU) has recommended one-way end-to-end delay no greater than 150 ms for good voice call quality, with a limit of 400 ms for acceptable voice calls [17].

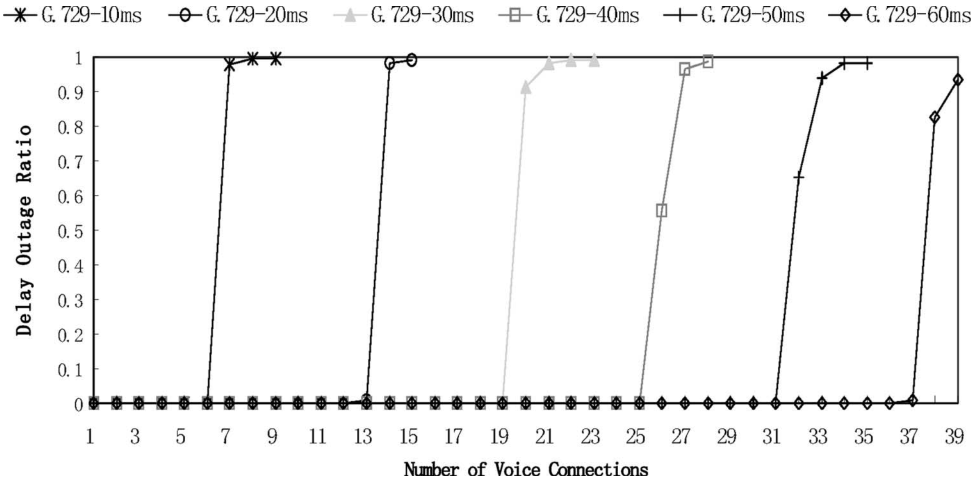

Fig. 7. Delay outage ratio of voice traffic.

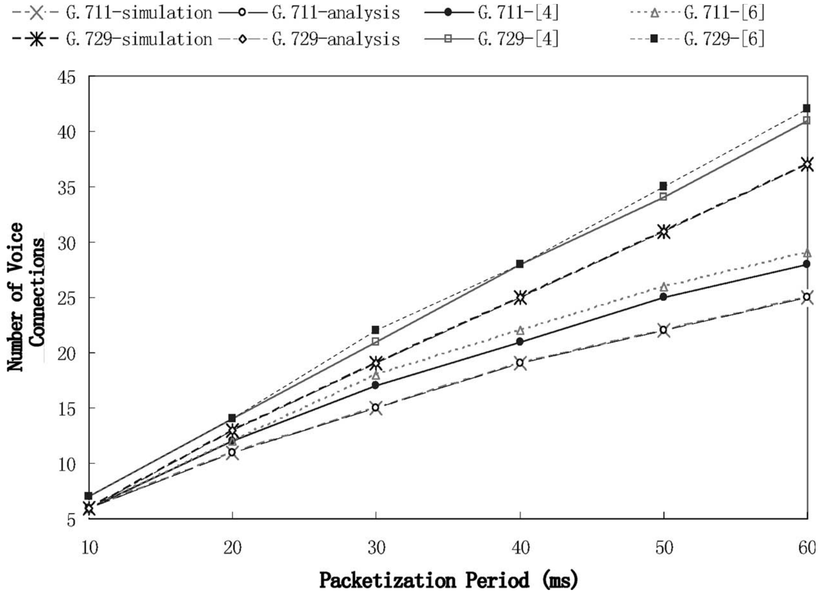

Fig. 8. Maximum number of voice connections.

our analytical results can be used as a guideline for admission control. Furthermore, from the simulation results, all frames are transmitted within five retransmissions, and none is dropped by the MAC due to excessive number of retransmissions, which validates our assumption in Section III.

VI. Conclusion

We have analytically studied the voice capacity of an infrastructure-based WLAN using the IEEE 802.11 DCF mode for MAC. It is concluded that the delay bound of real-time applications can be guaranteed only when the AP is not saturated. In addition, the number of voice connections that can be supported in an IEEE 802.11 WLAN is very limited due to the large overheads in each layer and the inherent inefficiency of the protocol. Our analytical results can be used as a guideline for effective admission control, which is necessary to guarantee the quality of voice traffic in WLANs. The performance of voice traffic in the presence of data traffic is under investigation.

References

[1] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai, "Investigation of the IEEE 802.11 medium access control (MAC) sublayer functions," in *Proc. IEEE INFOCOM*, Apr. 1997, vol. 13, pp. 126–133.

[2] M. Veeraraghavan, N. Cocker, and T. Moors, "Support of voice services in IEEE 802.11 wireless LANs," in *Proc. IEEE INFOCOM*, Anchorage, AK, Apr. 2001, vol. 1, pp. 448–497.

[3] Y. Xiao, "IEEE 802.11e: QoS provisioning at the MAC layer," *Wireless Commun.*, vol. 11, no. 3, pp. 72–79, Jun. 2004.

[4] F. Anjum *et al.*, "Voice performance in WLAN networks—An experimental study," in *Proc. IEEE GLOBECOM*, Dec. 2003, vol. 6, pp. 3504–3508.

[5] S. Garg and M. Kappes, "An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks," in *Proc. IEEE WCNC*, Mar. 2003, vol. 3, pp. 1748–1753.

[6] D. P. Hole and F. A. Tobagi, "Capacity of an IEEE 802.11b wireless LAN supporting VoIP," in *Proc. IEEE ICC*, Jun. 2004, vol. 1, pp. 196–201.

[7] S. Garg and M. Kappes, "Can I add a VoIP call?," in *Proc. IEEE ICC*, May 2003, vol. 2, pp. 779–783.

[8] 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.

[9] O. Tickoo and B. Sikdar, "Queuing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004, vol. 2, pp. 1404–1413.

[10] ——, "A queuing model for finite load IEEE 802.11 random access MAC," in *Proc. IEEE ICC*, Jun. 2004, vol. 1, pp. 175–179.

[11] UCB/LBNL/VINT, *Network Simulator NS-2 (2.27).* [Online]. Available: http://www.isi.edu/nsnam/ns/

[12] *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification,* Aug. 1999. Available: IEEE 802.11 WG

[13] S. Brunner and A. Ali, "Voice over IP 101-understanding VoIP networks," Juniper Networks, Sunnyvale, CA, Aug. 2004. Tech. Rep 200087-001.

[14] U. Black, *Voice Over IP*, 2rd ed. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[15] Global IP Sound, *iLBC—Designed for the Future*, Oct. 2004.

[16] A. Heck, *Introduction to Maple*, 3rd ed. New York: Springer-Verlag, 2003.

[17] International Telecommunication Union, *One-Way Transmission Time*, May 2003.

[18] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 3084–3094, Nov. 2005.

Lin X. Cai received the B.Sc. degree in computer science from Nanjing University of Science and Technology, Nanjing, China, in 1996 and the M.A.Sc. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2005. She is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Waterloo.

Her research interests include network performance analysis and protocol design for multimedia applications over wireless networks.

Xuemin (Sherman) Shen (M'97–SM'02) received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 1982 and the M.Sc. and Ph.D. degrees from Rutgers University, Piscataway, NJ, in 1987 and 1990, respectively, all in electrical engineering.

From September 1990 to September 1993, he was first with the Howard University, Washington, DC, and then with the University of Alberta, Edmonton, AB, Canada. Since October 1993, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, where he is a Professor and the Associate Chair for Graduate Studies. He is the author or coauthor of three books and more than 200 papers and book chapters in wireless communication and network control and filtering. His research interests focus on the mobility and resource management in interconnected wireless/wireline networks, ultra-wideband (UWB) wireless communications systems, wireless security, and *ad hoc* and sensor networks.

Dr. Shen serves as the Technical Program Committee Chair for Qshine'05, Co-Chair for IEEE Broadnet'05, WirelessCom'05, International Federation for Information Processing (IFIP) Networking'05, 2004 International Symposium on Parallel Architectures, Algorithms and Networks, and IEEE Globecom'03 Symposium on Next Generation Networks and Internet. He is also an Associate Editor of the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, the Association for Computing Machinery (ACM) *Wireless Network*, *Computer Networks*, *Dynamics of Continuous, Discrete and Impulsive Systems–Series B: Applications and Algorithms*, *Wireless Communications and Mobile Computing* (Wiley); and the *Computer Networks* (Elsevier). He has been Guest Editor of the IEEE Journal on Selected Areas in Communications (JSAC), *IEEE Wireless Communications*, and *IEEE Communications Magazine*. He received the Premier's Research Excellence Award (PREA) from the province of Ontario, Canada, for the demonstrated excellence of his scientific and academic contributions in 2003 and the Distinguished Performance Award from the Faculty of Engineering, University of Waterloo, for his outstanding contribution in teaching, scholarship, and service in 2002 and 2004. He is a registered Professional Engineer in Ontario.

Jon W. Mark (S'60–SM'80–F'88–LF'03) received the Ph.D. degree in electrical engineering from McMaster University, Hamilton, ON, Canada, in 1970.

In September 1970, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, where he served as the Department Chairman from July 1984 to June 1990 and is currently a Distinguished Professor Emeritus. In 1996, he established the Centre for Wireless Communications at the University of Waterloo and is currently serving as its Founding Director. He had been on sabbatical leave at the following institutions: IBM Thomas J. Watson Research Center, Yorktown Heights, NY, as a Visiting Research Scientist (1976–1977); AT&T Bell Laboratories, Murray Hill, NJ, as a Resident Consultant (1982–1983); Laboratoire MASI, Université Pierre et Marie Curie, Paris, France, as an Invited Professor (1990–1991); and the Department of Electrical Engineering, National University of Singapore, Singapore, as a Visiting Professor (1994–1995). He is a coauthor of the text *Wireless Communications and Networking* (Englewood Cliffs, NJ: Prentice-Hall, 2003). He has previously worked in the areas of adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, asynchronous transfer mode switch design, and traffic management. His current research interests are broadband wireless communications, resource and mobility management, and cross-domain interworking.

Dr. Mark was the recipient of the 2000 Canadian Award for Telecommunications Research and the 2000 Award of Merit of the Education Foundation of the Federation of Chinese Canadian Professionals. He was an Editor of the IEEE Transactions on Communications (1983–1990), a member of the Inter-Society Steering Committee of the IEEE/ACM Transactions on Networking (1992–2003), a member of the IEEE Communications Society Awards Committee (1995–1998), an Editor of Wireless Networks (1993–2004), and an Associate Editor of Telecommunication Systems (1994–2004).

Lin Cai (S'00–M'06) received the M.A.Sc. and Ph.D. degrees (with Outstanding Achievement in Graduate Studies Award) in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2002 and 2005, respectively.

Since July 2005, she has been an Assistant Professor in the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. Her research work has been published in prestigious journals and conferences (e.g., IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, IEEE Transactions on Multimedia, IEEE Transactions on Vehicular Technology, ACM MobiCom, and IEEE Infocom). She serves as an Associate Editor of the *EURASIP Journal on Wireless Communications* and *Networking and the International Journal of Sensor Networks*. Her research interests include wireless communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic over wireless, mobile, *ad hoc*, and sensor networks.

Yang Xiao (SM'04) received the B.Sc. degree in computational mathematics from Jilin University, Changchun, China, in 1989 and the M.Sc. and Ph.D. degrees in computer science and engineering from Wright State University, Dayton, OH, in 2000 and 2001, respectively.

He was with Micro Linear as a Medium Access Control Architect involved in the IEEE 802.11 standard enhancement work before he joined the Department of Computer Science, University of Memphis, Memphis, TN, in 2002. He was a Voting Member of the IEEE 802.11 Working Group from 2001 to 2004. He currently serves as the Editor-in-Chief of the *International Journal of Security and Networks* and the *International Journal of Sensor Networks*. He serves as an Associate Editor or on the Editorial Boards for the following refereed journals: the *International Journal of Communication Systems* (Wiley); *Wireless Communications and Mobile Computing* (Wiley); *EURASIP Journal on Wireless Communications and Networking*; and the *International Journal of Wireless and Mobile Computing*. He has served as a Lead/Sole Journal Guest Editor for five journals from 2004 to 2005. He serves as a referee for many funding agencies, as well as a panelist for the U.S. National Science Foundation. His research interests include wireless networks and network security.