Quality-of-Service Provisioning System for Multimedia Transmission in IEEE 802.11 Wireless LANs

Der-Jiunn Deng and Hsu-Chun Yen

***Abstract*—IEEE 802.11, the standard of wireless local area networks (WLANs), allows the coexistence of asynchronous and time-bounded traffic using the distributed coordination function (DCF) and point coordination function (PCF) modes of operations, respectively. In spite of its increasing popularity in real-world applications, the protocol suffers from the lack of any priority and access control policy to cope with various types of multimedia traffic, as well as user mobility. To expand support for applications with quality-of-service (QoS) requirements, the 802.11E task group was formed to enhance the original IEEE 802.11 medium access control (MAC) protocol. However, the problem of choosing the right set of MAC parameters and QoS mechanism to provide predictable QoS in IEEE 802.11 networks remains unsolved. In this paper, we propose a polling with non-preemptive priority-based access control scheme for the IEEE 802.11 protocol. Under such a scheme, modifying the DCF access method in the contention period supports multiple levels of priorities such that user handoff calls can be supported in wireless LANs. The proposed transmit-permission policy and adaptive bandwidth allocation scheme derive sufficient conditions such that all the time-bounded traffic sources satisfy their time constraints to provide various QoS guarantees in the contention free period, while maintaining efficient bandwidth utilization at the same time. In addition, our proposed scheme is provably optimal for voice traffic in that it gives minimum average waiting time for voice packets. In addition to theoretical analysis, simulations are conducted to evaluate the performance of the proposed scheme. As it turns out, our design indeed provides a good performance in the IEEE 802.11 WLAN's environment, and can be easily incorporated into the hybrid coordination function (HCF) access scheme in the IEEE 802.11e standard.**

***Index Terms*—Carrier sense multiple access/collision avoidance (CSMA/CA), point coordination function (PCF), quality-of-service (QoS), 802.11, wireless local area networks (WLAN).**

I. Introduction

FLEXIBILITY and mobility have made *wireless local area networks* (WLANs) a rapidly emerging field of activity in computer networking, attracting significant interests in the communities of academia and industry [1]–[6]. In the meantime, the IEEE standard for WLANs, IEEE 802.11 [1], has gained global acceptance and popularity in wireless computer networking markets and has also been anticipated to continue being the preferred standard for supporting WLAN's applications.

In WLANs, the medium access control (MAC) protocol is the key component that provides the efficiency in sharing the common radio channel, while satisfying the quality-of-service (QoS) requirements of various multimedia traffic. That is, MAC protocols that aim to carry multimedia traffic must be able to meet a variety of requirements for a wide range of traffic classes. However, frames in distributed coordination function (DCF), the basic access method in the IEEE 802.11 MAC layer protocol, do not have priorities, and there is no other mechanism to enforce a guaranteed access delay bound. As a result, real-time applications such as voice or live video transmissions may suffer from unacceptable delay with this protocol. The second access mode of the IEEE 802.11 MAC layer protocol, point coordination function (PCF), offers a "packet-switched connection-oriented" service, which is well suited for real-time traffic. However, in order to poll the stations an access point (AP) must maintain a polling list, which is implementation dependent. What this means is that end-to-end QoS requirements still cannot be satisfied in this scheme since it does not include any access control policy. Besides, it does not include any priority scheme to support handoff management, nor does it apply any bandwidth allocation strategy for handoff calls. However, packet-switched solutions, taking advantage of silences in a given voice call by multiplexing voice data from other calls, are more bandwidth-efficient than circuit-switched solutions. Furthermore, digitized multimedia traffic can be compressed to prevent tremendous bandwidth consuming comparing to uncompressed audio or video traffic. Since the demand for the use of packet-switched techniques for transferring delay-sensitive data in wireless environments is inevitable for multimedia applications, several works [7]–[31] have been investigated and discussed along this line of research. Accordingly, the IEEE 802.11 working group is currently working on a new standard called 802.11e [32] to enhance the original 802.11 MAC sublayer to support applications with QoS requirements.

Although the 802.11e standard is for providing QoS support for WLAN applications, the problem of choosing the right set of MAC parameters and QoS mechanism to provide predictable QoS in 802.11 networks remains unsolved [33]. Besides, the process of creating a definitive standard might be too slow for us waiting for it to be ratified. Hence, we implement a QoS provisioning system for multimedia transmission in IEEE 802.11 Wireless LANs, with a view to cutting, to a great extent, the telecommunication costs incurred to international enterprises.

Manuscript received January 31, 2004; revised December 1, 2004.

D.-J. Deng is with the Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan and also with the Department of Information Management, Overseas Chinese Institute of Technology, Taichung 407, Taiwan (e-mail: djdeng@cobra.ee.ntu.edu.tw).

H.-C. Yen is with the Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: yen@cc.ee.ntu.edu.tw).

Digital Object Identifier 10.1109/JSAC.2005.845632

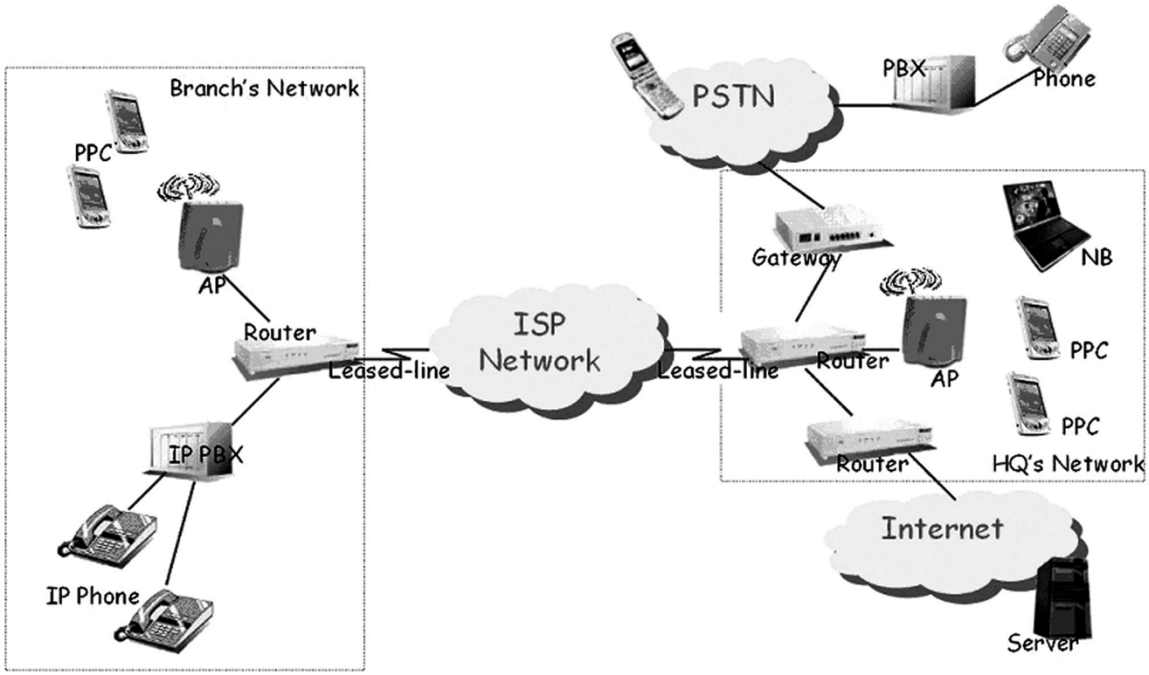

Fig. 1. System architecture.

In this paper, we propose an advanced, pragmatic, and yet more complete polling with nonpreemptive priority-based access control scheme for the IEEE 802.11 protocol. Our primary contributions are as follows: Under such a scheme, by modifying the DCF access method in the contention period, the protocol offers multiple levels of priorities such that handoff calls can be supported in wireless LANs. Besides, the proposed transmit-permission policy and adaptive bandwidth allocation scheme not only separate admitted inactivated users from newly requesting access users, but also derive sufficient conditions such that all the time-bounded traffic sources satisfy their time constraints to provide various QoS guarantees in the contention free period while maintaining efficient bandwidth utilization at the same time. Furthermore, the proposed scheduling algorithm for voice traffic is provably optimal in that it gives the minimum average waiting time for voice packets.

The proposed scheme is performed at each AP in a distributed manner. Such a scheme can be implemented in a broad class of algorithms with relatively minor modifications. In addition to theoretical analysis, simulations are conducted to evaluate the performance of the proposed scheme for integrated traffic. As it turns out, our design indeed provides good performance improvements over the original IEEE 802.11 protocol or the upcoming IEEE 802.11e standard.

The remainder of this paper is organized as follows. Section II describes the proposed scheme in detail. Simulation and experimental results are reported in Section III. Section IV concludes this paper.

II. Proposed Scheme

In this section, we present the proposed scheme in detail. Fig. 1 shows an overview of the proposed system architecture. Our method involves three basic components: 1) a *priority enforcement mechanism for request access*; 2) *a packet transmit-permission policy*;, and 3) *an adaptive bandwidth management strategy*.

A. Priority Enforcement Mechanism for Request Access

Since a mobile device travels while a connection is alive, the QoS might degrade because of some physical constraints. The problem will become even more challenging because recent wireless networks have been implemented using architecture based on small-size cells (i.e., microcells or picocells) to obtain higher transmission capacity and to achieve better performance. In most of the solutions, bandwidth is reserved for handoff mobiles in advance to reduce the dropping probability [35]. Some improvements have also been discussed [36]. However, when reserved and unused, the bandwidth is simply wasted. This is where priority schemes come in. The prioritized medium access of the EDCF in IEEE 802.11e is to provide service differentiation by allowing faster access to the channel to traffic classes with higher priority. Faster access can be provided by allocating a smaller contention window (CW) or a smaller interframe space (IFS). However, differentiating the initial CW size is better than differentiating the IFS in terms of total throughput and delay [27]. The reason is that the different initial CW size has both the function of reducing collisions and providing priorities, whereas the arbitration IFS has the function of providing priorities, but can not reduce collisions. In this section, we propose a novel method to modify the DCF access method to get many levels of priorities, capable of giving handoff requests higher priority over new connection requests. Under such a scheme, a high priority station is entitled to a shorter waiting time when accessing the medium. Furthermore, when a collision occurs, a high priority station can also take precedence over its lower priority counterparts in accessing the medium. Besides, the proposed

adaptive CW mechanism can dynamically expand and contract the CW size according to the current channel status. The method is simple, efficient, flexible, and scalable, based only on carrier sensing which can easily be implemented. It could be used as the random access protocol for contention period within a superframe in the IEEE 802.11 protocol without requiring any complicated computation or additional hardware support.

The core of the underlying collision avoidance mechanism in carrier sense multiple access with collision avoidance (CSMA/CA) is built upon a random backoff procedure, which generates a random backoff time (an integer value corresponding to the number of time slots). Initially, a station computes a backoff time in the range of 0–7. If a station with a frame to transmit initially senses a busy channel, it waits until the channel becomes idle, and then, the station decrements its backoff timer until either the medium becomes busy again or the timer reaches zero. If the medium becomes busy before the timer reaches zero, the station freezes its timer. When the timer finally decrements to zero, the station transmits its frame. If two or more stations decrement to zero at the same time, a collision occurs, and each station will have to generate a new backoff time in the range of 0–15. For each retransmission attempt, the backoff time grows in the form of $\lfloor \mathrm{ranf}(\,) \cdot 2^{2+i} \rfloor \cdot \mathrm{Slot\_Time}$, where i is the number of consecutive times a station attempts to send a frame, $\mathrm{ranf}(\,)$ is a uniform variate in (0, 1). (Here, $\lfloor x \rfloor$ represents the largest integer less than or equal to x.) For more about the CSMA/CA and the IEEE 802.11 protocols, the reader is referred to [34].

The basic idea behind our method is that prioritized access to the wireless medium is controlled through different backoff time periods. To this end, instead of using the one defined in [1], we change the backoff time generation function to $\lfloor \mathrm{ranf}(\,) \cdot (2^{2+i})/(2) \rfloor$ for high priority stations and $(2^{2+i})/(2) + \lfloor \mathrm{ranf}(\,) \cdot (2^{2+i})/(2) \rfloor$ for low priority stations. That is, the random backoff time is divided into two parts: $0 \sim (2^{2+i})/(2) - 1$ and $(2^{2+i})/(2) \sim 2^{2+i} - 1$, which are used by the high priority stations and the low priority ones, respectively. It is clear that the shorter backoff time a station waits, the higher priority this station will get. For example, initially (i.e., $i = 1$) the high priority stations generate a backoff time in the range of 0–3, and the low priority stations generate a backoff time in the range of 4–7. The former clearly has the edge on the latter in contending the channel. Such an idea can easily be generalized to support multiple-level priorities. Once again, the backoff time generation function is refined as $\lfloor \mathrm{ranf}(\,) \cdot 2^{1+i} \rfloor + k \cdot 2^{1+i}$, where k is the level of priority. Within a fixed backoff range, the probability of collisions with respect to the same priority level will increase if the number of contended stations increases. To offer a higher degree of flexibility and expandability, it is desirable that the scheme be able to expand or contract the backoff range arbitrarily. To be more precise, we allow different backoff ranges for different priority levels in our scheme by changing the backoff time generating function to $\lfloor \mathrm{ranf}(\,) \cdot 2^{m+i} \rfloor + k \cdot 2^{n+i}$, where k is the level of priority, and m and n are the parameters used to decide the number of slots in individual priority levels and the number of slots between each priority levels, respectively. In this paper, the real-time handoff traffic requests have the highest priority among all other requests, and the second priority class is the admitted inactivated video traffic. The new requests and pure data traffic will reside on the lowest priority level, as illustrated in Table I. Note that we give a wider range to the lowest priority level since this type of traffic is likely to be heavier, in comparison with the other two traffic classes. It is important to note that when a station decrements its backoff timer and the medium becomes busy, the station freezes its timer, meaning that a station will raise its priority automatically after several transmission failures. Hence, the proposed mechanism is clearly starvation-free.

TABLE I
EXAMPLES OF BACKOFF TIME OF INDIVIDUAL TRAFFIC

Consecutive times (i) / Backoff slot numbers / Types of requests (k, m, n)	1st	2nd	3 rd	4 th
Real-time handoff traffic (0, 1, 1)	0 – 3	0 -7	0 – 15	0 – 31
Admitted inactivated video traffic (1, 1, 1)	4 – 7	8 - 15	16 – 31	32 - 63
Non-real-time handoff traffic New request traffic (2, 2, 1)	8 – 15	16 - 31	32- 63	64 – 127

The collision avoidance portion of CSMA/CA is performed by a variable time-spreading of the users' access. However, collisions still occur if two or more stations select the same backoff slot. When this happens, these stations have to reenter the competition with an exponentially increasing CW parameter value, and the increase of the CW parameter value after collisions is the mechanism provided by CSMA/CA to make the access adaptive to channel conditions. This strategy avoids long access delays when the load is light because it selects an initial (small) parameter value of CW by assuming a low level of congestion in the system. However, it incurs a high collision probability and channel utilization is degraded in bursty arrival or congested scenarios. In other words, this strategy might allocate initial size of CW, only to find out later that it is not enough when the load increases. The size of CW must be reallocated with a larger size, but each increase of the CW parameter value is obtained paying the cost of a collision (bandwidth wastage). Furthermore, after a successful transmission, the size of CW is set again to the minimum value without maintaining any knowledge of the current channel status. Besides, the performance of CSMA/CA access method will be severely degraded not only in congested scenarios but also when the bit-error rate (BER) increases in the wireless channel. One principal problem also comes from the backoff algorithm. In the CSMA/CA access method, immediate positive acknowledgment informs the sender of successful reception of each data frame. This is accomplished by the receiver initiating the transmission of an acknowledgment frame after a small time interval, SIFS, immediately following the reception of the data frame. In case an acknowledgment is not received, as we mentioned above, the sender will presume that the data frame is lost due to collision, not by frame loss. Consequently, when a timer goes off, it exponentially increases backoff parameter value and retransmits the data frame less vigorously.

Unfortunately, wireless transmission links are noisy and highly unreliable. The proper approach to dealing with lost frames is to send them again, and as quickly as possible. Extending the backoff time just makes the matter worse because it brings bandwidth wastage.

As mentioned earlier, our scheme has the ability to expand or contract the backoff range arbitrarily by changing the parameters k, m and n. Hence, we propose an adaptive CW mechanism for our scheme to dynamically expand and contract the CW size according to the current channel status and achieve the theoretical capacity limits. This scheme is based on the results of the capacity analysis model of the IEEE 802.11 protocol originally proposed in [37]–[40], as well as the concept introduced in [41]. However, our scheme is simpler and more efficient and accurate, and it does not suffer from the problem of harmful fluctuation reported in [42].

In order to exploit the early and meaningful information about the actual congestion status of a channel, we start by defining the utilization factor α_c of a CW for real-time handoff traffic to be the number of busy slots s_c observed in the first to $2^{m_c+i}-1$ slots divided by the size (number of slots) of the current CW for admitted voice traffic (2^{m_c+i}) in the latest CW. The utilization factor, α_v, for admitted inactivated video traffic is defined by the number of busy slots s_v observed in the 2^{n_v+i} to $2^{m_v+i}+2^{n_v+i}-1$ slots divided by the size of the current CW for admitted inactivated video traffic (2^{m_v+i}) in the latest CW. The utilization factor α_a for new request and pure data traffic is defined to be the number of busy slots s_a observed in the 2^{n_a+i+1} to $2^{m_a+i}+2^{n_a+i+1}-1$ slots divided by the size of the current CW for new request and pure data traffic (2^{m_a+i}) in the latest CW. Note that the level of priority, k, for voice, video and new request/pure data traffic is 0, 1, and 2, respectively, here.

The arguments of how to adjust the CW size to achieve the theoretical capacity limits for these three classes of traffic are essentially the same. Hence, we assume that only one kind of traffic exists in the following argument. Note, however, that the parameters m and n of different traffic should be adjusted at the same time while changing any one of these parameters.

In practice, the value of α has to be updated in every backoff interval to reflect the actual state of the channel. Assume that there are M stations working in asymptotic conditions in the system. This means that the transmission queue of each station is assumed to be always nonempty. The stations transmit frames whose sizes are independent and identically distributed (i.i.d.) sampled from a geometric distribution with parameter q. Specifically, the size of a frame is an integer multiple of the slot size, t_{slot} and, hence, the mean frame space is $t_{\text{slot}}/1-q$.

Let $t_{\text{frame}}, t_{\text{virtual}}$, and t_{success} denote the average frame transmission time, the average temporal distance between two consecutive successful transmission, and the average time required for a successful transmission, respectively. Hence, the protocol capacity ρ is $t_{\text{frame}}/t_{\text{virtual}}$. Also, from the geometric backoff assumption, all the processes which define the occupancy pattern of the channel are regenerative with respect to the sequence of time instants corresponding to the completion of a successful transmission. Hence, the average time required for a successful transmission t_{succes} is bounded above by $t_{\text{frame}} + \text{ACK} + \text{DIFS} + \text{SIFS} + 2\cdot\tau$, where τ denotes the maximum propagation delay. Since an idle period is made up of a number of consecutive slots in which the transmission medium remains idle due to the backoff and the collisions and frame loss might occur between two consecutive successful transmissions, we have

$$t_{\text{virtual}} = E\left[\sum_{i=1}^{N}(\text{idel\_}p_i + \text{coll}_i + \text{lost}_i + \tau + \text{DIFS})\right] + E\left[\text{idel\_}p_{N_{\text{collision}}+N_{\text{lost}}+1}\right] + E[t_{\text{success}}] \quad (1)$$

where $\text{idel\_}p_i, \text{lost}_i$, and coll_i are the lengths of the i-th idle period, frame loss and collision in a virtual time, respectively, and $N_{\text{collision}}$ and N_{lost} are the number of collisions and the number of lost frames in a virtual time, respectively.

The assumption that the backoff interval is sampled from a geometric distribution with parameter p implies that the future behavior of a station does not depend on the past. Hence, the above equation can be rewritten as

$$t_{\text{virtual}} = E[N_{\text{collision}}]\cdot(E[\text{coll}] + \tau + \text{DIFS}) + E[N_{\text{lost}}]\cdot(E[\text{lost}] + \tau + \text{DIFS}) + E[\text{idle\_}p] \cdot (E[N_{\text{collision}}] + E[N_{\text{lost}}] + 1) + E[t_{\text{success}}]. \quad (2)$$

Closed expressions for $E[\text{idle\_}p], E[\text{lost}]$, and $E[\text{coll}]$ have been derived in the literature with $E[N_{\text{collision}}]$ and $E[N_{\text{lost}}]$

$$E[\text{idel\_}p] = \frac{(1-p)^M}{1-(1-p)^M}\cdot t_{\text{slot}} \quad (3)$$

$$E[N_c] = \frac{1-(1-p)^M}{M\cdot p\cdot(1-p)^{M-1}} - 1 \quad (4)$$

$$E[N_{\text{lost}}] = M\cdot\left(1-(1-p)^{M-1}\cdot(1-\text{BER})^{\frac{t_{\text{slot}}}{1-q}}\right) \quad (5)$$

$$E[\text{coll}] = E[\text{lost}] = \frac{t_{\text{slot}}}{1-[(1-p)^M + M\cdot p\cdot(1-p)^{M-1}]} \cdot\left[\sum_{h=1}^{\infty}(h\cdot((1-pq^h)^M - (1-pq^{h-1})^M)) - \frac{M\cdot p(1-p)^{M-1}}{1-q}\right]. \quad (6)$$

Hence, t_{virtual} is a function of the system's parameters, the number of active stations (M), the parameter p which defines the geometric-distribution used in the backoff algorithm, and the parameter q that characterizes the frame-size geometric distribution. As mentioned earlier, each station transmits a frame with probability p. This yields

$$P_{\text{error}} = 1-(1-p)^{M-1}\cdot(1-\text{BER})^{\frac{t_{\text{slot}}}{1-q}} \quad (7)$$

where p_{error} is the probability that a transmitted frame encounters a collision or is received in error, and BER denotes the channel bit error rate. Using the Markov chain, we can obtain an explicit expression for the probability p as a function of probability p_{error}

$$p = \frac{2(1-2p_{\text{error}})}{(1-2p_{\text{error}})(W+1) + W\cdot p_{\text{error}}(1-(2p_{\text{error}})^m)} \quad (8)$$

where W is the minimum CW, and m is the maximum number of backoff stages, i.e., $\text{CW} = W\cdot 2^m$. From (7), we obtain

$$M = 1 + \frac{\log\left(\frac{1-p_{\text{error}}}{(1-\text{BER})^{t_{\text{slot}}/1-q}}\right)}{\log(1-p)}. \quad (9)$$

Substituting p, as expressed by (8), into (9), we obtain

$$M = 1 + \frac{\log\left(\frac{1-p_{\text{error}}}{(1-\text{BER})^{t_{\text{slot}}/1-q}}\right)}{\log\left(1 + \frac{2\cdot(1-2\cdot p_{\text{error}})}{(2+W)\cdot p_{\text{error}} + W\cdot p_{\text{error}}(2\cdot p_{\text{error}})^m - (1+W)}\right)}. \quad (10)$$

Recall that the probability p_{error} is defined as the probability that a frame transmitted by the considered station fails. Since in each busy slot an eventual frame transmission would have failed, the probability p_{error} can be obtained by summing up the number of experienced collisions, frame losses, as well as the number of observed busy slots, and then dividing this sum by the total number of observed slots on which the measurement is taken, i.e., $p_{\text{error}} \approx \alpha$.

In order to maximize the utilization of every slot in a CW, we still need to engineer the tight upper bound of α to help us complete this scheme. We start with defining p_{opt} to be the value of p parameter that minimizes t_{virtual}. Since p_{opt} is closely approximated by the p value that guarantees a balance between the collision and frame loss and the idle periods in a virtual transmission time. Suppose there are M_{tr} stations making a transmission attempt in a slot. Then, we have

$$M \cdot p_{\text{opt}} = \sum_{i=1}^{M} i \cdot p\{M_{\text{tr}} = i\} \geq 1 - p\{M_{\text{tr}} = 0\} = \alpha. \quad (11)$$

As a consequence, $M \cdot p_{\text{opt}}$ is a tight upper bound of α in a system operating with the optimal channel utilization level. Substituting M, as expressed by (10), we obtain

$$p_{\text{opt}} \geq \frac{\alpha}{M} = \frac{\alpha}{1 + \frac{\log\left(\frac{1-\alpha}{(1-\text{BER})^{t_{\text{slot}}/1-q}}\right)}{\log\left(1+\frac{2\cdot(1-2\cdot\alpha)}{(2+W)\cdot\alpha+W\cdot\alpha\cdot(2\cdot\alpha)^m-(1+W)}\right)}}. \quad (12)$$

More precisely, the capacity of the 802.11 DCF protocol can be improved to achieve the theoretical throughput limit corresponding to the ongoing network environment, channel BER, and traffic configuration by dynamically adjusting its CW whose average size is identified by the optimal p value, p_{opt}, that is, when the average size of CW is $2/p_{\text{opt}} - 1$.

A natural strategy for expansion and contraction is to allocate a new CW size at the end of each transmission time. However, such a common heuristic would conduct the size of CW to fluctuate rapidly between expansion and contraction. To avoid this undesirable behavior, each station runs the algorithm to estimate the optimal CW size, and use the following formula to update its CW:

$$\text{New\_CW} = \chi\cdot\text{Current\_CW} + (1-\chi)\cdot\text{Estimate\_Optimal\_CW}$$

where $\chi \in [0,1]$ is a smoothing factor. Finally, instead of using the backoff time generation function defined in the IEEE 802.11 standard, we refine the backoff time generation function as $\lfloor \text{ranf}(\,) \cdot 2^{\lceil \log(\text{New\_CW}) \rceil} \rfloor \cdot t_{\text{slot}}$ to complete our scheme.

B. Packet Transmit-Permission Policy for Real-Time Traffic

Serving for the purpose of deciding whether a network accepts a new connection or not, the design of a packet transmission policy is one of the important challenges of traffic control in wireless networks. The policy is also used by the AP to determine which station gets permission to transmit a packet, especially in the realm of providing QoS. In this section, we present the methodology to highlight both admission control and reservation to meet the QoS requirements. Consequently, end-to-end QoS can be satisfied in WLANs. In what follows, we propose a packet transmit-permission policy for the IEEE 802.11 protocol to support integrated multimedia traffic. Our scheme is an enhanced version of the transmitting policy originally proposed in [43], and substantially extended from [44]. As we shall see later, our method is simple and efficient and can be implemented in the present IEEE 802.11 PCF or the upcoming IEEE 802.11e HCF standard easily. In addition, we also take handoff traffic into consideration. Under such a scheme, all voice traffic satisfies their jitter constraints, all video traffic satisfies their delay constraints, and the remaining bandwidth is shared by data traffic fairly and efficiently. Furthermore, the proposed scheduling algorithm for voice traffic is provably optimal, ensuring the minimum average waiting time for voice packets.

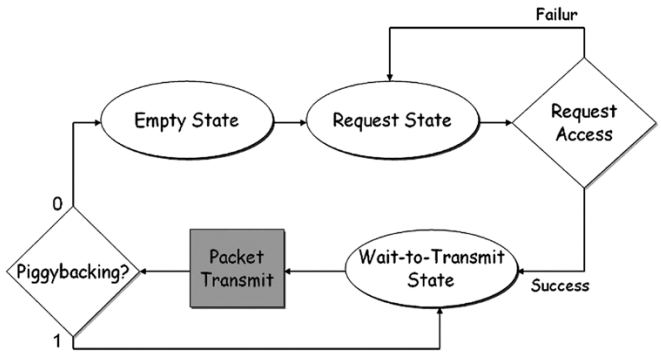

Fig. 2. State transition diagram of real-time stations.

At the IP layer, the maximum transmission unit (MTU) is set to be 1500 bytes, which is the maximum MSDU (the packet delivered to the MAC layer by the higher layer) size for the 100basedT Ethernet. Similarly, the 802.11 standard also provides a fragmentation mechanism, which allows the MAC layer to split an MSDU into more MPDUs (packets delivered by the MAC layer to the PHY layer). Hence, to formalize our problem, we assume that all real-time traffic packets have the same size in this paper. Besides, two types of real-time traffic are considered. The first is voice traffic which is characterized by two parameters (r_c, δ), where r_c is the rate (number of packets per second) of the source and δ is the maximum tolerable jitter (packet delay variation) for this stream. Jitter is defined to be the difference between the time of two successive departures and the time of two successive arrivals. The second is video traffic which is characterized by three parameters r_v, β and d, where r_v is the average rate of the source, β is the maximum burstiness of the source, and d is the maximum tolerable delay (packet transfer delay) for this stream.

The channel model considers the real-time stations to be in one of three states: "Empty," "Request," and "Wait to Transmit." Stations with empty buffers are said to be in the Empty state. When a packet (or packets) arrives to the buffer of a station in the Empty state, the station enters the Request state. A station in the Request state sends its request via the request access in

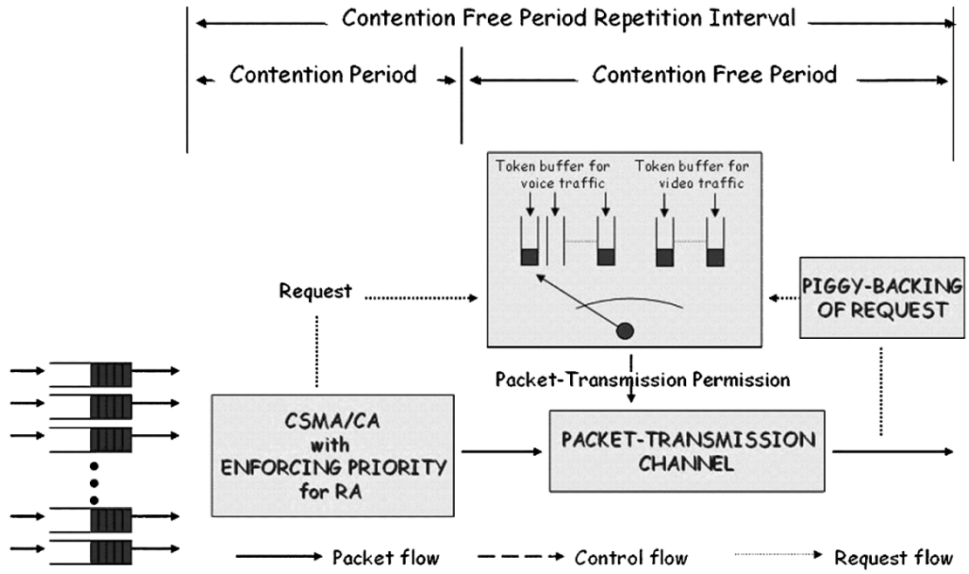

Fig. 3. Proposed packet transmit-permission policy.

contention period (CP) and stays in the Request state until its request is successfully received by the AP. When a station in the Request state successfully sends a request, it switches to the Wait-to-Transmit state. A station in the Wait-to-Transmit state listens to the channel until it is polled by AP, at which point it transmits a packet in the contention free period (CFP) of the next time slot, and also transmits a contention-free request (if necessary) using the PGBK request bit. If the station transmits a nonzero PGBK request bit (indicating the station's buffer is still nonempty), the station stays in the Wait-to-Transmit state. If the station transmits a zero PGBK request bit (indicating the station's buffer is empty), the station returns to the Empty state. Fig. 2 shows the state transition diagram of the real-time station.

In one basic service set (BSS) of the IEEE 802.11 infrastructure network architecture, the AP implements a token buffer for each real-time source. In token buffers for voice sources, the smaller the average rate is, the higher the priority becomes. In token buffers for video sources, the one with the smallest maximum delay constraint has the highest priority among all video sources. We depict the packet transmit-permission policy in Fig. 3. In order to gain control of the medium, the AP performs the function of the point coordinator by transmitting a beacon frame at the beginning of the CFP after sensing the medium to be idle for a PIFS period. Once the AP has the control of the medium, it performs the following tasks.

1) The AP first scans the token buffers of voice sources according to the preset priority order. If a token is found, it removes one from this token buffer and polls this voice terminal. On receiving a poll the station transmits its packet after a SIFS interval. Then, the AP generates the next token for this voice source after $(1)/(r_c) - (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + \mathrm{t_p} + \mathrm{ACK})$ second if the piggyback was set while transmitting the packet, where t_p is the time to transmit a real-time traffic packet.
2) If no tokens are found in the token buffers of voice sources, the AP continues to scan the token buffers for video sources according to the preset priority order. If a token is found, it polls this video source. And it will not remove the token if the piggyback was set while this video source transmitting it packet. If the piggyback was not set and it is not the last packet (end-of-file) either, the AP removes the token, and then generates the next token for this video source after η seconds if there is no new token generated for this video source within η, where η will be defined later.
3) If there is no token found in all token buffers, the AP will not know which, if any, of the stations has packets to transmit, then, it can end the CFP by transmitting a CF-end frame, and, for assuring the time constraint of admitted real-time traffic, the AP shall announce the beginning of the next CFP interval by observing the token buffer of highest priority on its polling list.

Note that the *contention-free multipoll* (CF-multipoll) mechanism defined in the IEEE 802.11e standard can be also used here to reduce the polling overhead from which the PCF suffers. In the following theorems, we provide sufficient conditions for all the voice packets to satisfy their maximum jitter constraints and for all the video packets to satisfy their maximum delay constraints, while optimizing the overall utilization of network bandwidth simultaneously.

Assume there are n_c voice sources (indexed by $i = 1, \ldots, n_c$), and n_v video sources (indexed by $j = 1, \ldots, n_v$). We denote (r_{ci}, δ_i) as the traffic parameters of the ith voice source, (r_{vj}, β_j, d_j) as the traffic parameters of the jth video source, and π_i as the time needed for handoff for source i.

Theorem 1: Let $\delta_1^* = 2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}$ and $\delta_i^* = 2 \cdot \mathrm{SIFS} + \mathrm{CFpoll} + t_p + \mathrm{ACK} + \sum_{k=1}^{i-1} \lceil (r_{ck})/(r_{ci}) \rceil \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK})$, $i = 2, \ldots, n_c$ and t_p be the time to transmit a packet. If $\delta_i^* < (1)/(r_{ci})$ and $\delta_i^* \leq \delta_i$ for all $i = 1, 2, \ldots, n_c$, then all the packets generated by new-call voice sources meet their jitter constraints. Furthermore, if $\delta_i^* + \pi_i < (1)/(r_{ci})$ and $\delta_i^* + \pi_i \leq \delta_i$ for the ith source which is

handoffed from other cells, then the packet generated by the ith source after handoff meets its jitter constraint. (Note that t_p should be adjusted when the bandwidth for real-time traffic changes. More precisely, t_p equals the size of MPDU divided by the bandwidth of channel I for real-time packets or the size of MPDU divided by the bandwidth of channel I and channel II for handoff real-time packets, where channel I and channel II will be defined later.)

Proof: We first prove the handoff part. Suppose that the first token generated from the ith voice source after handoff from other cells has a maximum waiting time $\bar{\delta}_i$. We want to prove that $\bar{\delta}_i \le \delta_i$ for $1 \le i \le n_c$. For $i = 1, \bar{\delta}_1 \le \pi_1 + 2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK} \le \pi_1 + \delta_1^* \le \delta_1$, which establishes the induction basis.

Suppose our induction hypothesis holds up to the $(i-1)$th voice sources, i.e., $\bar{\delta}_j \le \delta_j$ for $1 \le j \le i-1$. Now, we consider the ith voice source. Let the instant of the beginning of handoff be at time 0. Assume that $\bar{\delta}_i > \delta_i^* + \pi_i$. Then, it means that up to time $\delta_i^* + \pi_i$ the channel must be serving all the voice sources from 1 to $i-1$. Since the total amount of packets that can be served within $(0, \delta_i^* + \pi_i)$ for these $i-1$ voice sources is at most $\sum_{k=1}^{i-1} \lceil r_{ck} \cdot (\delta_i^* + \pi_i) \rceil$. Hence, the total amount of time to serve these packets is bounded above by $(\sum_{k=1}^{i-1} \lceil r_{ck} \cdot (\delta_i^* + \pi_i) \rceil + 1) \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK})$, and since $\delta_i^* + \pi_i < (1)/(r_{ci})$, we have

$$\begin{aligned}
&\left(\sum_{k=1}^{i-1} \lceil r_{ck} \cdot (\delta_i^* + \pi_i) \rceil + 1 \right) \\
&\quad \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \\
&\quad < \left(\sum_{k=1}^{i-1} \left\lceil \frac{r_{ck}}{r_{ci}} \right\rceil + 1 \right) \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \\
&\quad = \delta_i^* \le \delta_i^* + \pi_i.
\end{aligned}$$

This contradicts our assumption that $\bar{\delta}_i > \delta_i^* + \pi_i$. Hence, $\bar{\delta}_i \le \delta_i^* + \pi_i \le \delta_i$.

Based on the principle of induction, the statement of the theorem follows. Using a similar reasoning, it is reasonably easy to show that all the packets generated by new-call voice sources will meet their jitter constraints. Q.E.D.

Theorem 2: Suppose n_c voice sources are scheduled in the given priority order. The average waiting time is minimized for voice packets if $r_{ci} \le r_{cj}$ for all $i < j$.

Proof: The proof is done by contradiction. Assume that there exists a minimum average time schedule containing $r_{ci} > r_{cj}$ with $i < j$. Recall that the total waiting time within $(0, t)$ is

$$\begin{aligned}
&\sum_{k=1}^{n_c} t \cdot r_{ck} \left(\frac{1}{r_{ck}} \sum_{l=1}^{k-1} r_{cl} + 1 \right) \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \\
&= \sum_{k=1}^{n_c} t \cdot (n_c - k + 1) \cdot r_{ck} \cdot (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}).
\end{aligned}$$

Since $(n_c - i) > (n_c - j)$ and $r_{ci} > r_{cj}$, then this cannot be a minimum total waiting time schedule because we could get a shorter total waiting time schedule by simply exchanging the scheduling order of voice sources i and j—a contradiction. The theorem is, therefore, proven. Q.E.D.

Theorem 3: See the equation at the bottom of the page. If $\sum_{k=0}^{n_v} \bar{r}_{vk} \le 1$ and $d_j^* \le d_j$ for all j, then the delay constraints are satisfied for all the new-call video sources. Furthermore, if $d_j^* - \eta_j \le d_j - \pi_j$ for jth source which is handoffed from other cells, then the packet generated by the jth source after handoff meets its delay constraint.

Proof: Consider a nonnegative, left limited, and right continuous stochastic process $A \equiv \{a(t), t \ge 0\}$. Let $A(t_1, t_2) = \int_{t_1}^{t_2} a(t)\, dt$. We say that A is (β, r_v)-upper constrained if $A(s, t+s) \le r_v t + \beta$ for all $s, t \ge 0$. Similarly, A is (β, r_v)-lower constrained if $A(s, t+s) \ge r_v t + \beta$ for all $s, t \ge 0$. Since the number of departures in $(t_1, t_2]$ from a (β, r_v)-leaky bucket is bounded above by $\beta + \lceil r_v (t_2 - t_1) \rceil$, the departure process from a (β, r_v)-leaky bucket is $(\beta + 1, r_v)$-upper constrained.

Now, consider the first video source. Let $C_1 \equiv \{c_1(t), t \ge 0\}$ be the stochastic process that denotes the available bandwidth to the first video source at time t. If the channel is available to the first video source at time t, then $c_1(t) = 1$. Otherwise, $c_1(t) = 0$.

As mentioned above, the maximum number of packets from the n_c voice sources that can be served in $(t_1, t_2]$ is at most $\sum_{i=1}^{n_c} \lceil r_{ci}(t_2 - t_1) \rceil$. Hence, the bandwidth that is available to the first video source in $(t_1, t_2]$ is at least $t_2 - t_1 - (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot \{1 + \sum_{i=1}^{n_c} [r_{ci}(t_2 - t_1) + 1]\}$.

Thus, $C_1(t_1, t_2) \ge (1 - (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot \sum_{i=1}^{n_c} r_{ci}) \cdot (t_2 - t_1) - (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot (n_c + 1)$. That is, C_1 is $(\bar{\beta}_0, 1 - \bar{r}_{v0})$-lower constrained.

Let $A_1 \equiv \{a_1(t), t \ge 0\}$ be the amount of workload of video source 1 that arrives at the channel at time t. Since the

$$\begin{aligned}
\bar{\beta}_0 &= (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot (n_c + 1), \\
\bar{r}_{v0} &= (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot \sum_{i=1}^{n_c} r_{ci}, \\
\bar{\beta}_j &= (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot (\beta_j + 1), \\
\bar{r}_{vj} &= (2 \cdot \mathrm{SIFS} + \mathrm{CFPoll} + t_p + \mathrm{ACK}) \cdot r_{vj}, \quad d_1^* = \eta_1 + \frac{\bar{\beta}_0 + \bar{\beta}_1}{1 - \bar{r}_{vo}}, \\
\text{and } d_j^* &= \eta_j + \frac{\sum_{k=0}^{j} \bar{\beta}_k + (2 \cdot \mathrm{SIFS} + \mathrm{CFpoll} + t_p + \mathrm{ACK}) \cdot \sum_{k=1}^{j-1} (r_{vk} \cdot d_k^*)}{1 - \sum_{k=0}^{j-1} \bar{r}_{vk}}, \quad \text{where } j = 2, \ldots, n_v.
\end{aligned}$$

number of departures in $(t_1, t_2]$ from the first video traffic is $(\beta_1 + 1, r_{v1})$ -upper constrained. We have $A_1(t_1, t_2) \leq (2 \cdot \text{SIFS} + \text{CFPoll} + t_p + \text{ACK}) \cdot [r_{v1}(t_2 - t_1) + \beta_1 + 1]$. This shows A_1 is $(\bar{\beta}_1, t_p r_{v1})$-upper constrained.

Consider an instant after the last packet was sent (but not the EOF packet) by the first video source. Mark the instant as time 0. By letting $q_1(t)$ be the amount of backlogged workload from the first video source in the channel at time t, we have $q_1(0) = 0$. Since the next token for the first video source will be generated at time η_1 at the latest. We have $q_1(t) = A_1(0,t) - C_1(\eta_1, t)$. Note that the delay for an arrival at time t is bounded by the amount of time needed to deplete $q_1(t)$, and the time to deplete $q_1(t)$ is bounded by $\inf\{d \geq 0 : A_1(0,t) - C_1(\eta_1, t+d) \leq 0\}$. Maximizing over t, we have the following upper bound for the maximum delay:

$$d_1^* = \sup_t \inf\{d \geq 0 : A_1(0,t) - C_1(\eta_1, t+d) \leq 0\}$$

or

$$d_1^* = \sup_t \inf\{d \geq 0 : A_1(0,t) - C_1(\pi_1, t+d) \leq 0\}$$

(for the handoff traffic).

Since $\sum_{k=0}^{n_v} \bar{r}_{vk} \leq 1$, we have $\bar{r}_{vo} + (2 \cdot \text{SIFS} + \text{CFPoll} + t_p + \text{ACK}) r_{v1} \leq 1$. Applying the upper constraint for A_1 and the lower constraint for C_1, we have $d_1^* = \eta_1 + (\bar{\beta}_0 + \bar{\beta}_1)/(1 - \bar{r}_{vo})$, or $d_1^* = \pi_1 + (\bar{\beta}_0 + \bar{\beta}_1)/(1 - r_{vo})$. This completes the argument for the first video source.

The argument for the jth video source is essentially the same as that for the first video source. However, the lower constraint for the channel needs to be modified since the jth video source only uses the remaining channel after all the voice sources and the first $j-1$ video sources. Since the maximum delay of the kth video source is bounded above by $d_k^*, k = 1, \ldots, j-1$, the number of packets from the kth source that can be served in $(t_1, t_2]$ is bounded above by $\beta_k + [r_{vk}(t_2 - t_1 + d_k^*)]$. Hence, the amount of workload from the kth source that can be served in $(t_1, t_2]$ is bounded above by $[r_{vk}(t_2 - t_1) + \beta_k + 1 + r_{vk} d_k^*] \cdot (2 \cdot \text{SIFS} + \text{CFPoll} + t_p + \text{ACK})$. Parallel to the argument for the first video source, the maximum delay of the jth video source is bounded above by the first equation at the bottom of the page (or the second equation at the bottom of the page for the handoff traffic). Q.E.D.

Finally, we still need to engineer η_j to complete this scheme. In order to maximize the bandwidth utilization, one should have η_j as large as possible. The largest η_j can be obtained by solving $d_j^* = d_j$. However, larger η_j will lead to unsmooth video traffic. Therefore, we give a higher priority to the admitted inactivated video traffic in the contention period in order to compensate this shortcoming.

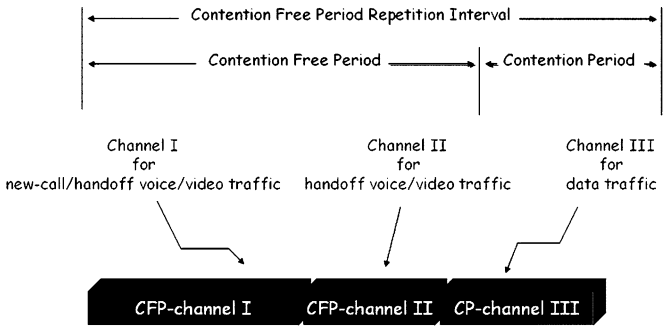

Fig. 4. Proposed bandwidth partition.

C. Adaptive Bandwidth Management Strategy

Recall that handoff calls can be supported in our proposed method since a real-time handoff station might be given higher priority over new connection requests. However, the handoff dropping probability might still increase if the desired amount of bandwidth in the neighboring cell, i.e., BSS, cannot be provided. In either case, binding the priority to channel access makes these QoS support mechanism unfair. As the number of stations generating high priority traffic increases, they tend to grab the channel. Hence, from the performance viewpoint, it is equally important to guarantee a minimum bandwidth for data traffic in order to maintain a reasonable bandwidth usage. To achieve this goal, in what follows we propose an adaptive bandwidth management strategy. Our strategy not only tries to maximize the bandwidth utilization and reduce the handoff dropping probability and blocking probability but also guarantees a minimum bandwidth for data traffic. In addition, this strategy is very simple and easy to implement without excessive computations.

As shown in Fig. 4, the total bandwidth is divided into three parts: channels I, II, and channel III. We allocate channel I for real-time traffic and channel II for handoff real-time traffic in contention free period. By allowing the handoff real-time traffic to use bandwidth exclusively with preemptive priority over other traffic in channel II, the handoff real-time traffic might have a larger share of pie in bandwidth utilization to reduce the dropping probability. Likewise, the remaining real-time traffic has precedence over the new request/data traffic for using network resources in channel I. Channel III is only reserved for new requests and data traffic to guarantee a minimum

$$\eta_j + \frac{\sum_{k=0}^{j} \bar{\beta}_k + (2 \cdot \text{SIFS} + \text{CFPoll} + t_p + \text{ACK}) \cdot \sum_{k=1}^{j-1} (r_{vk} \cdot d_k^*)}{1 - \sum_{k=0}^{j-1} \bar{r}_{vk}}$$

$$\pi_j + \frac{\sum_{k=0}^{j} \bar{\beta}_k + (2 \cdot \text{SIFS} + \text{CFPoll} + t_p + \text{ACK}) \cdot \sum_{k=1}^{j-1} (r_{vk} \cdot d_k^*)}{1 - \sum_{k=0}^{j-1} \bar{r}_{vk}}$$

bandwidth for data traffic in the contention period. However, after bandwidth is allocated, network conditions may change. Therefore, the proposed strategy can also adjust the amount of allocated bandwidth based on the measured dropping probability, blocking probability, and bandwidth utilization. The algorithm to control the size of the allocated bandwidth is summarized in the following.

Function Adaptive_Bandwidth_Allocation
IF monitored dropping probability $>$ threshold_D **THEN**
  IF bandwidth utilization $< \mu$ **THEN**
    size of allocated bandwidth II = min {max {size of allocated bandwidth I, size of allocated bandwidth II} $\times$ up_γ, total bandwidth }
  ELSE
    size of allocated bandwidth II = min {max {size of allocated bandwidth I, size of allocated bandwidth II} $\times$ up_γ, total bandwidth $\times$ threshold_channel_II_max }
ELSE
  IF monitored blocking probability $>$ threshold_B **THEN**
    IF bandwidth utilization $< \mu$ **THEN**
      size of allocated bandwidth I = min {size of allocated bandwidth I $\times$ up_γ, total bandwidth $\times$ threshold_channel_I_max }
    ELSE
      size of allocated bandwidth I = min {size of allocated bandwidth I $\times$ up_γ, total bandwidth $\times$ threshold_channel_I_medium }
  ELSE
    IF bandwidth utilization $< \mu$ **THEN**
      size of allocated bandwidth II = max {size of allocated bandwidth II $\times$ down_γ, total bandwidth $\times$ threshold_channel_II_min }
      size of allocated bandwidth I = max {size of allocated bandwidth I $\times$ down_γ, total bandwidth $\times$ threshold_channel_I_min }

As the pseudocode illustrates, the handoff dropping probability is the first measure used to adjust the allocated bandwidth. If the dropping probability over the threshold, threshold_D, and the bandwidth utilization is not good enough (less than the threshold value μ), it implies that there is not so much data traffic. Hence, we increase the size of channel II by a factor up_γ (by changing the value of t_p, i.e., the total time of transmits a real-time packet in a contention free period, in the inequalities of Theorems 1 and 3) to its maximum (total bandwidth). Otherwise, we guarantee a minimum bandwidth for data traffic by only increasing the size of channel II to the threshold (total bandwidth $\times$ threshold_up_II). Then, we use the blocking probability to adjust the allocated bandwidth of channel I in the same way. However, to lower dropping probability will get higher priority than to lower blocking probability in adjusting bandwidth allocation. Finally, the allocated bandwidth will be stable in a good situation if the bandwidth is over the threshold μ. That is, both dropping probability and blocking probability are under the threshold, and the bandwidth utilization is above the threshold value μ.

III. Performance Evaluation

In this section, we evaluate the performance of the proposed scheme.

A. Simulation Environment

Our simulation model is built using the Simscript tool [45]. The model represents a BSS in the IEEE 802.11 WLANs with all stations in the BSS capable of directly communicating with the remaining parties. To focus on the access control issue and to reduce the complexity of the simulation, what follows are the basic assumptions in our simulation environment. First, the "hidden terminal" and the "exposed terminal" problems [46] are not addressed in the simulation model. Second, no stations operate in the "power-saving" mode. Third, no interference is considered from nearby BSSs. Finally, the probability of a frame being transmitted successfully is calculated as: $p_r\{\text{success}\} = (1 - \text{BER})^n$, where n is the number of bits transmitted in the frame and BER denotes the bit error rate.

Three types of traffic are considered in the simulation.

1) Pure data: The arrival of data frames from a station's higher layer to MAC sublayer is Possion. Frame length is assumed to be exponentially distributed with mean length 1024 octets.
2) Voice traffic: Voice stream is characterize by two parameters (γ_c, δ), where γ_c is the rate of the source and δ is the maximum tolerable jitter (packet delay variation) for this stream. Frames of voice traffic that are not successfully transmitted within its maximum jitter constraint are assumed to be lost. The voice stream is modeled as a two state Markov on/off process, where stations are either transmitting (on) or listening (off). The amount of time in the off or on state is exponentially distributed, where the mean value of the silence (off) period is 1.5 s, and the mean value of the talk spurt (on) period is 1.35 s. The duration of each connection is exponentially distributed with mean time 3 min.
3) Video traffic: Each video stream is characterized by three parameters (r_v, β, d), where r_v is the average rate of the source, β is the maximum burstiness of the source, and d is the maximum tolerable delay (packet transfer delay) for this stream. We use a source model in [47]. The bit rate of a single source for the nth frame $\lambda(n)$ is defined by the recursive relation: $\lambda(n) = a\lambda(n-1) + bw(n)$ [bit/pixel], where $a = 0.8781, b = 0.1108$, and $w(n)$ is a sequence of independent Gaussian random variables which have mean 0.572 and variance 1. Like voice frames, video frames that are not successfully transmitted within its maximum tolerable delay, d, are assumed to be lost.

The default values used in the simulation are listed in Table II. The values for the simulation parameters are chosen carefully in order to closely reflect the realistic scenarios, as well as to make the simulation feasible and reasonable.

B. Numerical and Simulation Results

In the first instance, we measure the maximum jitter and delay for the real-time sources both through the analytical models and

TABLE II
DEFAULT ATTRIBUTE VALUES USED IN THE SIMULATION

Attribute	Value	Meaning & Explanation
Channel rate	11 Mb/s	Data rate for the wireless channel
Stations	20	20 mobile hosts in a basic service set
Slot_Time	20 *us*	Time needed for each time slot
SIFS	10 *us*	Time needed for each short interframe space
PIFS	30 *us*	Time needed for each PCF interframe space
DIFS	50 *us*	Time needed for each DCF interframe space
MAC header	272 bits	Header length of MAC layer header
PHY header	192 bits	Header length of physical layer header
ACK	112 bits + PHY header	Frame length of each Acknowledgement
Beacon	432bits + PHY header	Frame length of each Beacon frame
CF-End	160 bits + PHY header	Frame length of each CF_end frame
BER	10^{-6}	Bit error rate
χ	0.5	Smoothing factor
μ	0.8	Minimum bandwidth utilization wanted
r_c	32 kb/s	Voice source data rate
δ	32 *ms*	Tolerable jitter for voice source
β	5	Maximum burstiness
π	5 *ms*	Time needed for handoff
d	50 *ms*	Maximum packet delay for video source
Threshold_D	0.1	Maximum allowable dropping probability
Threshold_B	0.1	Maximum allowable blocking probability
up_ r	1.1	Bandwidth allocation each time increases 10%
Down_ r	0.9	Bandwidth allocation each time decreases 10%
threshold_channel_II_max	0.9	Channel II uses at most 90% of total bandwidth
threshold_channel_I_max	0.8	Channel I uses at most 80% of total bandwidth when blocking rate is too large and bandwidth utilization too low
threshold_channel_I_mid	0.7	Channel I uses at most 70% of total bandwidth when blocking rate is too large and bandwidth utilization is high enough
threshold_channel_II_min	0.5	Channel II uses at least 50% of total bandwidth
threshold_channel_I_min	0.3	Channel I uses at lease 30% of total bandwidth
Handoff probability	0.25	The probability that a mobile moves out of the range of a base station

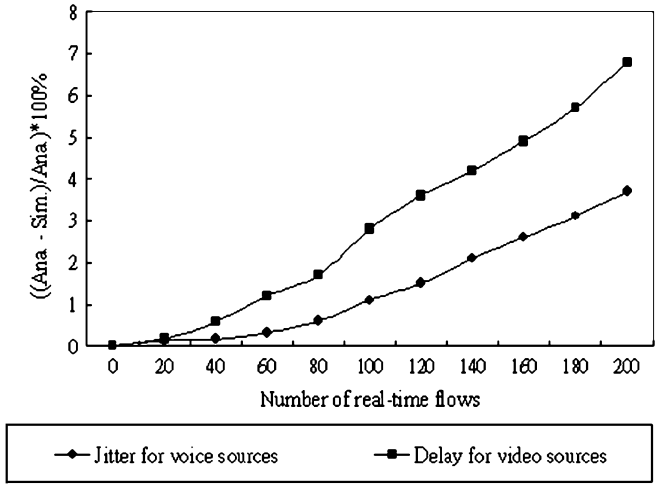

Fig. 5. Differences between analytical models and simulated results.

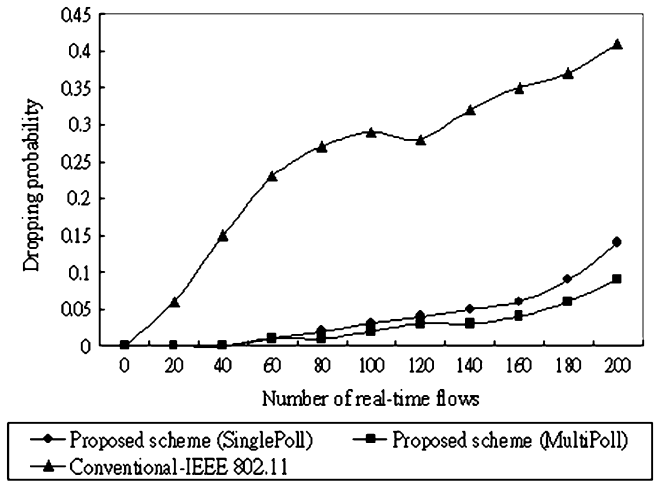

Fig. 6. Dropping probability of real-time handoff connections.

simulations to verify the correctness and usefulness of the proposed scheme. In Fig. 5, we show the differences between the analytical models and the simulated results for voice and video sources. They both show that the restricted bounds of maximum jitter/delay in the proposed scheme are more conservative as they are derived from the worst case analysis. Although both of the figures show that the results are quite close, the simulated results still have about 3% to 7% inaccuracy.

In what follows, the performances of the proposed scheme and the conventional IEEE 802.11 protocol are compared based on simulations. The performances of CF-multipoll mechanism are also discussed. In the CF-multipoll mechanism, the AP can poll more than one station simultaneously using a single CFPoll frame to reduce the polling overhead. In the conventional IEEE 802.11 protocol, CSMA/CA is adopted as the random access protocol for the contention period, and a round-robin discipline is chosen as the scheduling policy for AP in the contention free period. That is, all traffics have the same priority. The admission control scheme in the conventional IEEE 802.11 protocol is very simple and intuitive. Assuming there are totally k requests in the request table, if $k\cdot(2\cdot\mathrm{SIFS}+\mathrm{CFPoll}+t_p+\mathrm{ACK})\le\delta_i$ for the ith voice source or $k\cdot(2\cdot\mathrm{SIFS}+\mathrm{CFPoll}+t_p+\mathrm{ACK})\le d_j$ for the jth video source, the request of a new voice or video source is admitted; otherwise, it is rejected. Understandably, the time needed for handoff π will be added for the handoff mobile in admission control. The duration of the contention free period and the length of each superframe are set to be 50 and 75 ms, respectively, for the conventional IEEE 802.11 protocol since the video, voice and data traffic are assumed to be mixed in the ratio of 1:1:1 and the maximum delay for video traffic is 50 ms in the simulation. The contention free period ends when either its duration has reached the maximum value, i.e., 50 ms, or the AP has no more requests in its request table. However, it is noteworthy that the duration of the contention free period and the length of each superframe are dynamically allocated according to the current channel and traffic status by the AP using the proposed packet transmit-permission policy in the proposed scheme.

Simulation results are shown below in the form of plots. Figs. 6 and 7 show the dropping probability of real-time handoff connections and blocking probability of real-time new connections for the proposed scheme and the conventional IEEE 802.11 protocol. Since the handoff dropping probability is the first measure used to adjust the allocated bandwidth

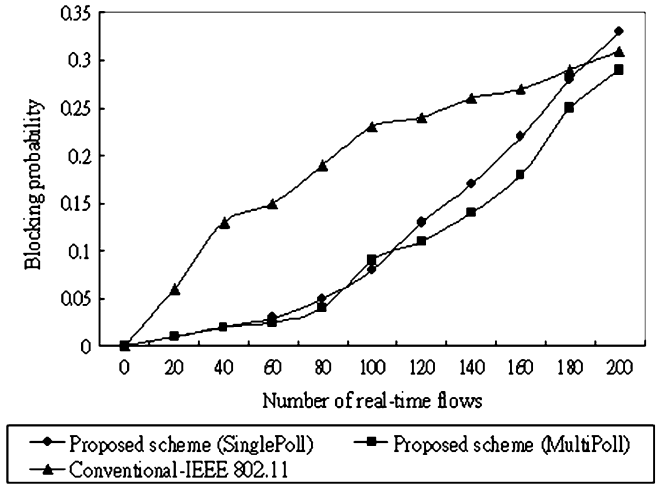

Fig. 7. Blocking probability of real-time new connections.

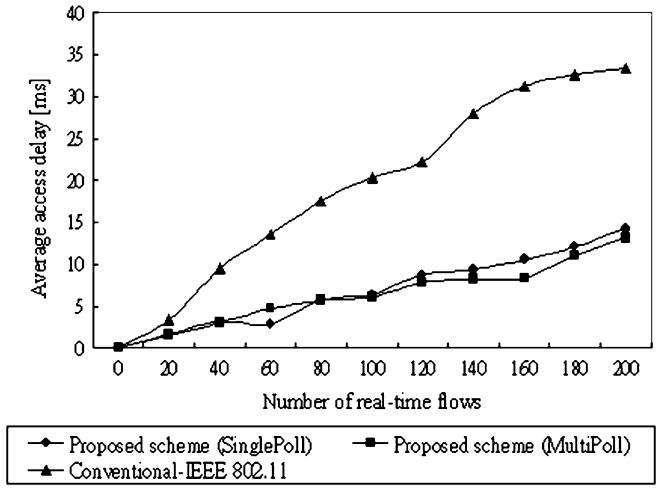

Fig. 8. Average access delay of voice traffic, variance = 21.17 (proposed scheme-single poll), 15.49 (proposed scheme-multipoll), and 136.09 (conventional IEEE 802.11).

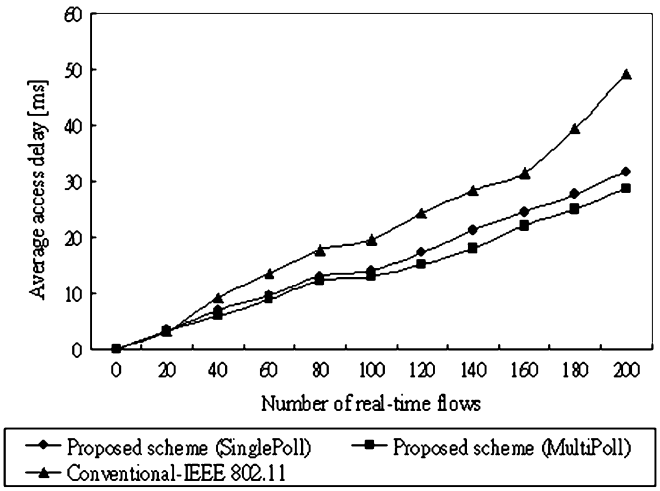

Fig. 9. Average access delay of video traffic, variance = 103.27 (proposed scheme-single poll), 82.73 (proposed scheme-multipoll), and 152.97 (conventional IEEE 802.11).

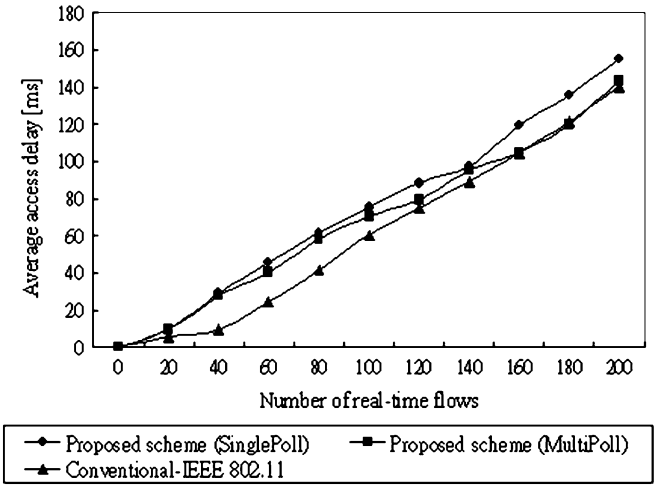

Fig. 10. Average access delay of data traffic, variance = 2577.15 (proposed scheme-single poll), 2116.55 (proposed scheme-multipoll), and 2397.36 (conventional IEEE 802.11).

and we also allow the handoff real-time traffic to use bandwidth exclusively with preemptive priority over other traffics in the reserved region, channel II, the dropping probability will normally be kept under the threshold, i.e., threshold_D. The fact that the proposed scheme provides a slightly higher blocking probability in heavy load is obvious because it shows the tradeoff between the dropping probability and the blocking probability. However, it seems counterintuitive that the proposed scheme provides lower blocking probability in light load. This is because that the contention free period can start as soon as the request table just becomes nonempty in the proposed scheme, then the AP will end the current contention period. However, once the contention period starts, the real-time traffic is not allowed to be served until the next contention free period in the conventional IEEE 802.11 protocol.

Figs. 8 and 9 compare the average access delays of voice and video traffic from the proposed scheme and the conventional IEEE 802.11 protocol, respectively. We can see that although there is not much difference in the values of the performance measures when load is light, however, the proposed scheme provides significantly better performance than the conventional IEEE 802.11 protocol at heavy load. In other words, the average access delay of the proposed scheme remains low when the offered load is high; in contrast, the conventional IEEE 802.11 protocol shows a sharp rise as the load increases since it lacks any priority/reservation scheme and access control policy. The simulation results suggest that the proposed scheme is appropriate for transmitting high priority real-time traffic such as voice and video traffic in real-time applications.

Fig. 10 shows the average access delay of data traffic in a multimedia communication environment. As expected, the average access delay of data traffic in the proposed scheme is worse than the conventional IEEE 802.11 protocol since it is of low priority. However, the lower priority traffic can have the bandwidth it needs in light load, so it is not wasted. In fact, a minimum bandwidth for data traffic to maintain a reasonable bandwidth usage can still be guaranteed by using the proposed adaptive bandwidth management scheme.

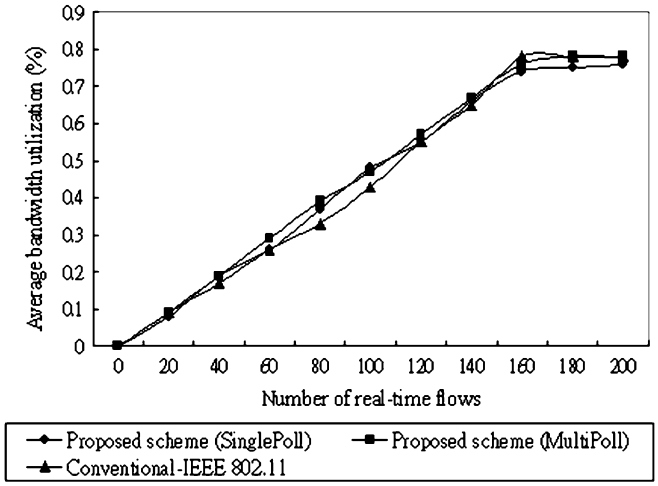

Fig. 11. Average bandwidth utilization.

Fig. 11 presents the average bandwidth utilization as a function of the offered load. Average bandwidth utilization is the percentage of the bandwidth actually being used in the total bandwidth. As illustrated in Fig. 11, the average bandwidth utilization is slightly lower for the proposed scheme in a highly loaded system because to maintain the desired QoS, it must be more conservative in admitting new connections. It reveals that there is a clear tradeoff between deterministic (hard) QoS supporting and bandwidth utilization. However, it also reveals that our proposed scheme reduces the handoff dropping probability without sacrificing the bandwidth utilization too much. Besides, the multipoll scheme can surely increase the overall performance over the single-poll scheme, as can easily be seen in the figure.

IV. Conclusion

In the era of multimedia communication, the design of priority-sensitive network protocols continues to be an important issue, and broadband wireless links constitute a subclass in which prioritization is key to optimizing the overall performance of the network. In this paper, we have proposed a pragmatic polling with nonpreemptive priority-based access control scheme built on well-known protocols, offering easily implemented and yet flexible criteria for traffic prioritization in a wireless environment. By modifying the DCF access method in the contention period, our designed protocol supports multiple levels of priorities such that user mobility can be supported in wireless LANs. Besides, the proposed transmit-permission policy and adaptive bandwidth allocation scheme not only separate admitted inactivated users from newly requesting access users, but also derive sufficient conditions such that all the time-bounded traffic sources satisfy their time constraints to provide various deterministic QoS guarantees in the contention free period while maintaining efficient bandwidth utilization at the same time. Furthermore, the proposed scheduling algorithm for voice traffic is provably optimal in that it gives minimum average waiting time for voice packets. The proposed scheme is performed at each AP in a distributed manner. Through extensive simulations, we have demonstrated a satisfactory performance of our proposed scheme in a quantitative way.

Acknowledgment

The authors would like to thank the anonymous referees for their helpful comments that greatly improved the quality of this paper.

References

[1] *IEEE Draft Standard P802.11*, Jan. 1996. Wireless Medium Access Control and Physical Layer WG, Wireless LAN, IEEE Std. Dept, D3.

[2] ETSI TC-RES, "Radio equipment and systems (RES); High performance radio local area network (HIPERLAN); Functional specification," ETSI, 06 921 Sophia Antipolis Cedex, France, draft prETS 300 652, Jul. 1995.

[3] M. Chelouche, S. Hethuin, and L. Ramel, "Digital wireless broadband corporate and private network: RENT concepts and applications," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 42–51, Jan. 1997.

[4] K. Y. Eng, M. J. Karol, M. Veeraraghavan, E. Ayanoglu, C. B. Woodworth, P. Pancha, and R. A. Valenzuela, "BAHAMA: A broadband ad hoc wireless ATM local-area network," in *Proc. ICC*, 1995, pp. 1216–1223.

[5] P. Pilat, "HomeRF-SWAP: Optimized for home networking," *Intel Developer Update Mag.*, pp. 1–5, Jun. 2000.

[6] B. A. Miller and C. Bisdikian, *Bluetooth Revealed—The Insider's Guide to an Open Specification for Global Wireless Communications*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[7] J. Geier, *Wireless LANs—Implementing High Performance IEEE 802.11 Network*, 2nd ed: Sams Publishing, Jul. 2001, p. 19.

[8] R. O. LaMaire *et al.*, "Wireless LAN's and mobile networking: Standards and future directions," *IEEE Commun. Mag.*, vol. 34, no. 8, pp. 86–94, Aug. 1996.

[9] R. A. Dayem, *Moble Data and Wireless LAN's Technology*. Englewood Cliffs, NJ: Prentice-Hall, 1997, pp. 190–201.

[10] M. Veeraraghavan, N. Cocker, and T. Moors, "Support of voice services in IEEE 802.11 wireless LANs," in *Proc. INFOCOM*, 2001, pp. 488–497.

[11] Y. Wang and B. Bensaou, "Priority based multiple access for service differentiation in wireless ad hoc networks," in *Proc. MWCN*, 2000, pp. 14–30.

[12] P. H. Chuang, H. K. Wu, and M. K. Liao, "Dynamic QoS allocation for multimedia ad hoc wireless networks," *Proc. Comput. Commun. Netw.*, pp. 480–485, Oct. 1999.

[13] J. L. Sobrinho and A. S. Krishnakumar, "Quality-of-service in ad hoc carrier sense multiple access wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 8, pp. 1353–1368, Aug. 1999.

[14] ——, "Distributed multiple access procedures to provide voice communications over IEEE 802.11 wireless networks," in *Proc. GLOBECOM*, 1996, pp. 1689–1694.

[15] D. J. Deng and R. S. Chang, "A priority scheme for IEEE 802.11 DCF access method," *IEICE Trans. Commun.*, vol. E82-B, no. 1, pp. 96–102, Jan. 1999.

[16] C. R. Lin, "Multimedia transport in multihop wireless networks," *Proc. Inst. Elect. Eng. Commun.*, vol. 145, no. 5, pp. 342–345, Oct. 1998.

[17] J. Y. Yen and C. H. Chen, "Support of multimedia services with the IEEE 802.11 MAC protocol," in *Proc. ICC* , vol. 1, Apr. 2002, pp. 600–604.

[18] S. T. Sheu and T. F. Sheu, "A bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LAN's," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2065–2080, Oct. 2001.

[19] S. Chakrabarti and A. Mishra, "QoS issues in ad hoc wireless networks," *IEEE Commun. Mag.*, vol. 39, pp. 142–148, Feb. 2001.

[20] A. Lindgren, A. Almquist, and O. Schelen, "Evaluation of quality of service schemes for IEEE 802.11 wireless LANs," in *Proc. LCN*, Nov. 2001, pp. 348–351.

[21] T. Suzuki and S. Tasaka, "Performance evaluation of priority-based multimedia transmission with the PCF in an IEEE 802.11 standard wireless LAN," in *Proc. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, vol. 2, Oct. 2001, pp. G70–G77.

[22] ——, "Performance evaluation of integrated video and data transmission with the IEEE 802.11 standard MAC protocol," in *Proc. GLOBECOM*, vol. 1b, Nov. 1999, pp. 580–586.

[23] S. Sharma, K. Gopalan, and N. Zhu, "Quality-of-service guarantee on 802.11 networks," *Proc. Hot Interconnects 9*, pp. 99–103, Aug. 2001.

[24] C. Coutras, S. Gupta, and N. B. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs," *Wireless Netw.*, vol. 6, no. 6, pp. 457–466, Dec. 2000.

[25] C. Andren, "IEEE 802.11 wireless LAN: Can we use it for multimedia?," *IEEE Multimedia*, vol. 5, no. 2, pp. 84–89, Apr. 1998.
[26] F. Eshghi and A. K. Elhakeem, "Performance analysis of ad hoc wireless LAN's for real-time traffic," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 204–215, Feb. 2003.
[27] Y. Xiao, "Enhanced DCF of IEEE 802.11e to support QoS," in *Proc. WCNC 2003*, Mar. 2003, pp. 1291–1296.
[28] S. C. Lo, G. L. Lee, and W. T. Chen, "An efficient multipolling mechanism for IEEE 802.11 wireless LANs," *IEEE Trans. Comput.*, vol. 52, no. 6, pp. 764–778, Jun. 2003.
[29] G. Anastasi and L. Lenzini, "QoS provided by the IEEE 802.11 wireless LAN to advanced data applications: A simulation analysis," *Wireless Netw.*, vol. 6, no. 2, pp. 99–108, Mar. 2000.
[30] O. Sharon and E. Altman, "An efficient polling MAC for wireless LAN's," *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 439–451, Aug. 2001.
[31] D. Gu and J. Zhang, "QoS enhancement in IEEE 802.11 wireless local area networks," *IEEE Commun. Mag.*, vol. 41, no. 6, pp. 120–124, Jun. 2003.
[32] *Part 11: Wireless Medium Access Control and Physical Layer Specifications: Medium Access Control Enhancements for Quality of Service*, Nov. 2002. IEEE 802.11e draft/D4.0.
[33] W. P. Atikom, P. Krishnamurthy, and S. B. Jee, "Distributed mechanisms for quality of service in wireless LANs," *IEEE Commun. Mag.*, vol. 10, no. 3, pp. 26–34, Jun. 2003.
[34] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai, "IEEE 802.11 wireless local area networks," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 116–126, Sep. 1997.
[35] B. Moon, C. Oh, A. Ahmad, and K. Kim, "A study of bandwidth allocation strategies in multimedia wireless networks," in *Proc. Conf. Record APCC*, 1997, pp. 509–513.
[36] C. Oliveira, J. Bae, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 6, pp. 858–873, Aug. 1998.
[37] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 wireless LAN: Capacity analysis and protocol enhancement," in *Proc. INFOCOM*, San Francisco, CA, Mar. 1998, pp. 142–149.
[38] L. Bononi, M. Conti, and E. Gregori, "Design and performance evaluation of an asymptotically optimal backoff algorithm for IEEE 802.11 wireless LANs," in *Proc. 33rd Hawaii Int. Conf. Syst. Sci.*, 2000, pp. 1–10.
[39] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. Netw.*, vol. 8, no. 6, pp. 785–799, Dec. 2000.
[40] 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.
[41] L. Alcuri, G. Bianchi, and I. Tinnirello, "Occupancy estimation in the IEEE 802.11 distributed coordination function," presented at the ICS2002, Hualien, Taiwan, 2003.
[42] L. Bononi, M. Conti, and L. Donatiello, "Design and performance evaluation of a distributed contention control (DCC) mechanism for IEEE 802.11 wireless local area networks," in *Proc. Workshop WOWMOM, MOBICOM*, Dallas, TX, Oct. 1998, pp. 1–10.
[43] D. J. Deng and R. S. Chang, "A nonpreemptive priority based access control scheme for broadband ad hoc wireless ATM local area networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 9, pp. 1731–1739, Sep. 2000.
[44] C. S. Chang, K. C. Chen, M. Y. You, and J. F. Chang, "Guaranteed quality-of-service wireless access to ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 1, pp. 106–118, Jan. 1997.
[45] CACI Products Company. (1997, Sep.). Simscript II.5, CA 92 037. [Online]. Available: http://www.caciasl.com/
[46] A. S. Tanenbaum, *Computer Networks*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996, pp. 263–264.
[47] B. Maglaris *et al.*, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834–844, Jul. 1998.

Der-Jiunn Deng, photograph and biography not available at the time of publication.

Hsu-Chun Yen, photograph and biography not available at the time of publication.