Achieving Optimal Performance by Using the IEEE 802.11 MAC Protocol With Service Differentiation Enhancements

Bo Li, Roberto Battiti, *Associate Member, IEEE*, and Yong Fang

***Abstract*—Wireless local area networks are currently evolving to support adequate degrees of service differentiation. Work is in progress to define an enhanced version of the IEEE 802.11 distributed coordination function, which is capable of supporting quality of service for multimedia traffic at the medium access control layer. In this paper, we aim at gaining insight into one mechanism to differentiate among traffic categories, i.e., differentiating the minimum contention window sizes according to the priority of different traffic categories. The contribution of this paper is the analysis of the optimal operation point where maximum throughput can be achieved. Through the analysis, we answer some fundamental questions about the existence and uniqueness of the optimal operation point, about the maximum system throughput, about the existence of simple rules to decide if the system operates under the optimal state or not, and about procedures to lead the system to the optimal operation point. The other contribution is the proposal of a simple adaptive scheme that makes the system operate under the optimal operation point and, at the same time, achieve target service differentiation between different traffic flows. The results that are obtained in this paper are relevant to both theoretical research and implementations of real systems.**

***Index Terms*—IEEE 802.11 medium access control (MAC), performance analysis, service differentiation, wireless local area network (WLAN).**

I. Introduction

THE MAIN objective of next-generation broadband wireless networks is to provide seamless multimedia services to mobile users. In this context, one of the major challenges of wireless mobile Internet is to provide suitable levels of quality of service (QoS) over Internet-Protocol-based wireless access networks [1]–[3]. Wireless access should be considered as just another hop in the communication path for the whole Internet. A good example for such a wireless technology is the IEEE 802.11 distributed coordination function (DCF) standard [4], which is compatible with the current best effort service model of the Internet. In the literature, performance evaluation of 802.11 has been executed by using a simulation [5] or by means of analytical models [6]–[11]. Constant or geometrically distributed backoff window sizes have been considered in [6]–[8]. In [9], an exponential backoff with only two stages is modeled by using a 2-D Markov chain. In [10], a more general model that accounts for all the exponential backoff protocol details is proposed. In [11], instead of using stochastic analysis, the average value for a variable is used, which results in an approximate but effective analysis.

In order to support different QoS requirements for various types of service, a possibility is to support service differentiation in the IEEE 802.11 medium access control (MAC) layer, as proposed in [12]–[15]. With service-differentiation support, it means that different types of traffic flows can obtain different QoSs, such as bandwidth, packet delays, and delay jitters. In [12], a simple priority scheme for IEEE 802.11 has been proposed, where a high-priority station has shorter waiting time when accessing the medium. In [13], a service-differentiation scheme is proposed. The scheme uses two parameters of the IEEE 802.11 MAC, namely: 1) the backoff interval and 2) the interframe space (IFS) between each data transmission, to provide the differentiation. In [14], service differentiation is supported by setting different minimum contention windows (CWs) for different types of services. Reference [15] proposes three service-differentiation schemes for the IEEE 802.11 DCF. The first one is based on scaling the CW according to the priority of each flow. The second one assigns different IFSs to different traffic classes. The third one uses different maximum frame lengths. Moreover, an effective CW resetting scheme to enhance the performance of the IEEE 802.11 DCF is analyzed in [16] by extending the model proposed in [10]. In [17], both the enhanced DCF (EDCF) and the hybrid coordination function, which are defined in the IEEE 802.11e draft [18], are evaluated through simulation.

In order to gain a deeper insight into the modified IEEE 802.11 MAC with service-differentiation support, system modeling and performance analysis are needed. In [19], a performance study of the IEEE 802.11 MAC protocol with service differentiation has been made. However, the model is complex, which makes it difficult to obtain deeper insight into the system performance. Li and Battiti [20], [21] propose a simple analysis model to compute the throughput in a wireless

Manuscript received September 16, 2004; revised July 8, 2005, June 5, 2006, and June 16, 2006. This work was supported in part by the National Science Foundation of China under Grant 60572144 and Grant 2005F27 funded by the Shaanxi Province, by the Chinese Government Program for New Century Excellent Talents in University under Grant NCET-06-0876, by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, and by the Wireless Internet and Location Management Architecture Project funded by the Provincia Autonoma di Trento, Italy. The review of this paper was coordinated by Prof. T. Hou.

B. Li and Y. Fang are with the School of Electronics and Information Engineering, Northwestern Polytechnic University, Xi'an 710072, China (e-mail: libo.npu@nwpu.edu.cn; libo1998@gmail.com; yfang79@gmail.com).

R. Battiti is with the Department of Computer Science and Telecommunications, University of Trento, 38050 Trento, Italy (e-mail: battiti@dit.unitn.it).

Digital Object Identifier 10.1109/TVT.2007.895565

local area network with enhanced IEEE 802.11 DCF. In the proposed analytical model, service differentiation is supported by differentiating the minimum CW size and the IFS according to the priority of each traffic flow. The proposed model is so simple that it can be solved analytically, which further makes it possible for one to obtain deeper insight into the system performance. Similar works on the performance analysis of the backoff-based priority schemes for IEEE 802.11 and IEEE 802.11e can also be found in [22] and [23].

Moreover, some more practical adaptive schemes are proposed to make the system cope with dynamic traffic better. In [24], a scheme of dynamically tuning the IEEE 802.11 protocol has been proposed. By using the scheme, maximum throughput can be achieved. However, in this paper, how to achieve service differentiation is not considered. In [25], an adaptive EDCF scheme is given. The scheme uses the idea of "slow decrease of CW size" to improve the utilization of the system. Service differentiation is considered in this paper. In [26], EDCF with dual measurement is proposed. The proposed scheme is based on the basic idea of reducing the number of idle slots and adapting a CW size according to the current traffic state and network conditions. Performance comparisons show that better QoS can be achieved. However, no rigorous analytical model is proposed on how to achieve maximum throughput and target service differentiation at the same time.

In this paper, based on our former work described in [20] and [21], we successfully analyze the optimal operation point where maximum saturation throughput can be achieved. It is worth noting that we confine our discussions only for the case where equal IFSs are adopted for different traffic flows. Moreover, a simple adaptive scheme that makes the system operate under the optimal operation point and achieve target service differentiation between different traffic flows is also proposed.

II. IEEE 802.11 DCF

In the IEEE 802.11 DCF protocol, when the MAC receives a request to transmit a frame, a check is made of the physical and virtual carrier sense mechanisms. If the medium is not in use for an interval of distributed IFS (DIFS), the MAC may begin transmission of the frame. If the medium is in use during the DIFS interval, the MAC selects a backoff time, which is randomly and uniformly chosen in the range of $(0, W-1)$, with W being the CW. The MAC decrements the backoff value each time the medium is detected to be idle for an interval of one slot time. The terminal starts transmitting a packet when the backoff value reaches zero. When a station transmits a packet, it must receive an acknowledgment (ACK) frame from the receiver after a short IFS (SIFS), or it will consider the transmission to have failed. If a failure happens, the station reschedules the packet transmission according to the given backoff rules and increments the retry counter. If there is a collision, the CW is doubled, and a new backoff interval is selected. At the first transmission attempt, W is set equal to a value $\mathrm{CW}_{\min}$, which is called the minimum CW. After each unsuccessful transmission, W is doubled, up to a maximum value $\mathrm{CW}_{\max} = 2^m \cdot \mathrm{CW}_{\min}$.

Furthermore, in order to overcome the hidden station problem, 802.11 defines an optional request-to-send/clear-to-send (RTS/CTS) mechanism. In this paper, only the *basic access* mechanism is analyzed. The analysis method can be easily extended to the case of *RTS/CTS access* mechanism.

The basic DCF method is not appropriate for handling multimedia traffic requiring guarantees about throughput and delay. Because of this weakness, task group E of the IEEE 802.11 working group is currently working on an enhanced version of the standard called IEEE 802.11e. The goal of the extension is to provide a distributed access mechanism that is capable of service differentiation. A new access mechanism called EDCF has been selected [27]. It is shown by simulation that EDCF has better performance than point coordination function (PCF) and is more scalable [28]. Note that the terminology "enhanced distributed channel access (EDCA)" is used instead of EDCF in the newer version of the IEEE 802.11e draft [18]. In the IEEE 802.11e EDCA, four backoff entities are supported within one 802.11e station, with each backoff entity corresponding to a particular access category (AC). Service differentiation for each AC is supported by using AC-specific contention parameters, which comprise the so-called EDCA parameter set. Arbitration IFS (AIFS [AC]) and the minimum CW size $\mathrm{CW}_{\min}[\mathrm{AC}]$ are included in the EDCA parameter set for each AC. With shorter AIFS[AC] and/or smaller $\mathrm{CW}_{\min}[\mathrm{AC}]$, the corresponding backoff entity in an 802.11e station has higher priority in accessing channel resources, which brings about relatively better QoS for the corresponding traffic flows.

In this paper, in the interest of conciseness, we are not interested in exploring all the details of the new proposed IEEE 802.11e standard but in gaining insight into one of the building blocks that are used to achieve service differentiation, i.e., differentiating the minimum CW sizes according to the priority of each traffic category. Moreover, for simplicity, only one backoff entity is supported in a sending station.

III. Performance Analysis

To make this paper self-contained, in this section, we briefly summarize the model and analysis that are presented in [20] and [21], dealing with extensions of DCF to support service differentiation. All details about the analysis can be found in [20].

A. System Modeling

We assume that the channel conditions are ideal (i.e., no hidden terminals and capture, and no transmission errors) and the system operates in saturation: A fixed number of traffic flows always have packets available for transmission.

In the current basic service set (BSS), $M(\geq 1)$ types of traffic are considered, with n_i type-$i(1 \leq i \leq M)$ traffic flows existing in the system. It is assumed that each station bears only one traffic flow. Let $b_i(t)$ be the stochastic process representing the backoff-time counter for a given type-$i(1 \leq i \leq M)$ traffic flow. Moreover, let us define for convenience $W_i \equiv \mathrm{CW}_{\min,i}$ as the minimum CW for type-i traffic flows. Let m_i be the "maximum backoff stage" such that $\mathrm{CW}_{\max,i} = 2^{m_i} W_i$. Let $s_i(t)$ be the stochastic process representing the backoff stage $(0, 1, \ldots, m_i)$ for a given type-i traffic flow.

The key approximation in the model is that, at each transmission attempt for a type-i traffic flow, regardless of the number of

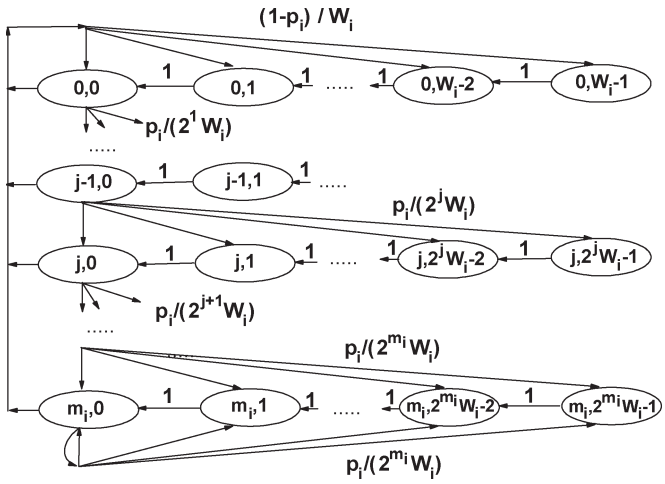


Fig. 1. Markov model of the backoff process for a type-i traffic flow.

retransmissions suffered, each packet collides with constant and independent probability p_i. This assumption has been shown by simulation to be accurate if W_i and n_i are large [10]. Parameter p_i is referred to as conditional collision probability, i.e., the probability of a collision that is seen by a packet to belong to a type-i traffic flow at the time of its transmission on the channel.

In the following, we use a 2-D discrete-time Markov chain to model the behavior of a type-i traffic flow. The states are defined as the combinations of two integers $\{s_i(t), b_i(t)\}$. The Markov chain for type-i traffic flows is shown in Fig. 1.

B. Throughput Analysis

Let $q_i(j,k)$, $j \in [0, m_i]$, and $k \in [0, 2^j W_i - 1]$ be the stationary distribution of the chain. τ_i is defined as the probability that a station carrying type-i traffic transmits in a randomly chosen slot time. We have

$$\tau_i = \sum_{j=0}^{m_i} q_i(j,0) = \frac{2(1-2p_i)}{(1-2p_i)(W_i+1) + p_i W_i \left[1-(2p_i)^{m_i}\right]}. \tag{1}$$

Detailed derivations on (1) can be found in [10]. With the preceding probabilities defined, we can express packet collision probabilities p_i as

$$p_i = 1 - (1-\tau_i)^{n_i - 1} \prod_{j=1, j\neq i}^{M} (1-\tau_j)^{n_j}. \tag{2}$$

After combining (1) and (2) and using successive over-relaxation [29], we can get all the values for p_i and τ_i.

In order to derive the system throughput, we define $Q(c_1, \ldots, c_M)$ as the probability that there are c_i $(0 \le c_i \le n_i), (1 \le i \le M)$ type-i stations transmitting within a randomly selected slot. Then, we have

$$Q(c_1, \ldots, c_M) = \prod_{i=1}^{M} \binom{n_i}{c_i} \tau_i^{c_i} (1-\tau_i)^{n_i - c_i}. \tag{3}$$

It is evident that if $\sum_{i=1}^{M} c_i \ge 2$, it corresponds to the case where more than one station transmits in the same selected time slot, which brings about packet collisions.

The normalized system throughputs S can be defined and expressed as (4), shown at the bottom of the page, where S_i denotes the total throughputs that are contributed by type-i traffic. $T_{\mathrm{Len},i}$ is the average time duration to transmit the payload for type-i traffic (the corresponding packet payload length, which is measured in bits, is denoted as $P_{\mathrm{Len},i}$). For simplicity, it is assumed that all packets of type-i traffic have the same fixed size. σ is the duration of an empty time slot. $T_{s,i}$ is

$$\begin{aligned}
S &\equiv \frac{\text{Average payload transmitted in a slot time}}{\text{Average length of a slot time}} \\
&= \sum_{i=1}^{M} S_i \\
&= \frac{\sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1\le j\le M, j\neq i)} = 0\right) \cdot T_{\mathrm{Len},i}}{\left\{ \begin{array}{c} Q\left(c_{j(1\le j\le M)} = 0\right) \cdot \sigma + \sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1\le j\le M, j\neq i)} = 0\right) \cdot T_{s,i} \\ + \sum_{0\le c_j \le n_j (1\le j\le M),\ \sum_{j=1}^{M} c_j \ge 2} Q(c_1, \ldots, c_M) \cdot T_c(c_1, \ldots, c_M) \end{array} \right\}} \\
&\equiv \frac{\sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1\le j\le M, j\neq i)} = 0\right) \cdot T_{\mathrm{Len},i}}{\left\{ \begin{array}{l} Q\left(c_{j(1\le j\le M)} = 0\right) \cdot \sigma + \sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1\le j\le M, j\neq i)} = 0\right) \cdot T_{s,i} \\ + \left[1 - Q\left(c_{j(1\le j\le M)} = 0\right) - \sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1\le j\le M, j\neq i)} = 0\right)\right] \cdot T_c \end{array} \right\}}
\end{aligned} \tag{4}$$

the average time of a slot because of a successful transmission of a packet of type-i traffic flow. $T_{s,i}$ can be expressed as

$$T_{s,i} = \text{PHY}_{\text{header}} + \text{MAC}_{\text{header}} + T_{\text{Len},i} + \text{SIFS} + \delta + \text{ACK} + \text{DIFS} + \delta \quad (5)$$

where δ is the propagation delay. $T_c(c_1, \ldots, c_M)$ is the average time that the channel is sensed to be busy by each station during a collision that is caused by simultaneous transmissions of $c_i (0 \le c_i \le n_i), (1 \le i \le M)$ type-i stations. It can be expressed as

$$T_c(c_1, \ldots, c_M) = \text{PHY}_{\text{header}} + \text{MAC}_{\text{header}} + \max\left[\theta(c_1)T_{\text{Len},1}, \ldots, \theta(c_M)T_{\text{Len},M}\right] + \text{DIFS} + \delta \quad (6)$$

where

$$\theta(x) \equiv \begin{cases} 1, & x > 0 \\ 0, & x = 0. \end{cases}$$

In (4), T_c is defined as the average duration that the channel is sensed busy during a collision. It can be explicitly given as (7), shown at the bottom of the page.

C. Approximation Analysis

From (2), we can easily derive

$$(1 - p_i)(1 - \tau_i) = \prod_{j=1}^{M} (1 - \tau_j)^{n_j}, \qquad 1 \le i \le M. \quad (8)$$

When the minimum CW size $W_i \gg 1$, transmission probabilities $\tau_i \ll 1$. Therefore, from (8), we have the following approximation:

$$p_i \approx p_j (i \ne j). \quad (9)$$

Furthermore, when $W_i \gg 1$ and $m_i \approx m_j (i \ne j)$, based on (1), the following approximation holds:

$$\tau_i W_i \approx \tau_j W_j, \qquad i \ne j. \quad (10)$$

From (3), (4), and (10), we have

$$\frac{s_i}{s_j} \equiv \frac{S_i/n_i}{S_j/n_j} = \frac{\frac{\tau_i}{1-\tau_i} \cdot T_{\text{Len},i}}{\frac{\tau_j}{1-\tau_j} \cdot T_{\text{Len},j}} \approx \left(\frac{T_{\text{Len},i}}{W_i}\right) \Big/ \left(\frac{T_{\text{Len},j}}{W_j}\right) \quad (11)$$

where $s_i (\equiv S_i/n_i) (1 \le i \le M)$ is defined as the average throughput that is contributed by an individual type-i traffic flow. It can be regarded as the bandwidth that is occupied by a sending station bearing a type-i traffic flow.

IV. Maximum Throughput Analysis

In this section, by using the previous analysis, we analyze the optimal operation point where the maximum throughput can be achieved. We are interested in maximizing total throughput S while *at the same time* ensuring service differentiation, and the hypothesis in this section is that service differentiation is achieved by allocating the bandwidth to individual traffic flows to satisfy a given target ratio $\hat{\alpha}_i \equiv s_i/s_1 \ (\hat{\alpha}_i > 0, 1 \le i \le M)$. Moreover, we define $\alpha_i \equiv ((\tau_i/(1-\tau_i))/(\tau_1/(1-\tau_1))) (1 \le i \le M)$, which is one part of s_i/s_1 [see (11)], as the ratio of the packet sending rates between a type-i and a type-1 traffic flow. In the following performance analysis, it can be seen that α_i takes an important role. According to (11), target bandwidth allocation ratio $\hat{\alpha}_i$ can be expressed as $\hat{\alpha}_i = \alpha_i \cdot T_{\text{Len},i}/T_{\text{Len},1} (1 \le i \le M)$, which indicates that $\hat{\alpha}_i$ is determined by both the ratio of packet sending rates α_i and the ratio of channel holding times $T_{\text{Len},i}/T_{\text{Len},1}$ between different types of traffic flows. In IEEE 802.11e, the duration of the channel holding times for packet transmission can be controlled by configuring proper transmission opportunity [18]. In this paper, in order to achieve the maximum throughput and target bandwidth allocation ratio $\hat{\alpha}_i$, we pay attention to the mechanism of controlling the ratio of packet sending rates α_i by adjusting the minimum CW sizes $W_i (1 \le i \le M)$ for each traffic flow.

In the following, we always assume that the packet sending rates for all the active traffic flows in the current BSS satisfy the constraint that $0 < \tau_i < 1 (1 \le i \le M)$.

Theorem 1: Assume that $M (\ge 1)$ types of traffic coexist in the system, with $n_i (1 \le i \le M)$ numbers of type-i traffic flows. In the case that $\tau_i/(1-\tau_i) = \alpha_i \cdot (\tau_1/(1-\tau_1)) (\alpha_i > 0, 1 \le i \le M, \alpha_1 \equiv 1)$, throughput function $S(\tau_1, \ldots, \tau_M)$, which is defined in (4), has one and only one optimal operation point $\tau_i^*(\alpha_1, \ldots, \alpha_M) (1 \le i \le M)$ where the maximum throughput can be achieved.

Proof: Based on (3) and (4), we can write the throughput function as (12) shown at the bottom of the next page, where $\chi \equiv \tau_1/(1-\tau_1)$, $F_1 \equiv \sum_{i=1}^{M} n_i \cdot \alpha_i \cdot T_{\text{Len},i}$, $G_1 \equiv \sum_{i=1}^{M} n_i \cdot \alpha_i \cdot T_{s,i}$, and

$$G_i \equiv \sum_{0 \le c_j \le n_j (1 \le j \le M), \sum_{l=1}^{M} c_l = i} T_c(c_1, \ldots, c_M) \cdot \prod_{k=1}^{M} \alpha_k^{c_k} \binom{n_k}{c_k} \left(2 \le i \le \sum_{j=1}^{M} n_j\right).$$

Moreover, F_1 and $G_i (1 \le i \le \sum_{j=1}^{M} n_j)$ are constants that are larger than zero. In order to determine the optimal operation

$$T_c \equiv \frac{\sum_{0 \le c_j \le n_j (1 \le j \le M), \ \sum_{j=1}^{M} c_j \ge 2} Q(c_1, \ldots, c_M) \cdot T_c(c_1, \ldots, c_M)}{1 - Q\left(c_{j(1 \le j \le M)} = 0\right) - \sum_{i=1}^{M} Q\left(c_i = 1, c_{j(1 \le j \le M, j \ne i)} = 0\right)} \quad (7)$$

point, we study function $(F(\chi)/G(\chi))'$, which can be given as

$$\left(\frac{F(\chi)}{G(\chi)}\right)' = \left(F_1\sigma - F_1\sum_{i=2}^{\sum_{j=1}^{M} n_j}(i-1)G_i\chi^i\right)\Big/ G(\chi)^2. \quad (13)$$

The optimal solution χ^* should satisfy the following equation:

$$\sum_{i=2}^{\sum_{j=1}^{M} n_j}(i-1)G_i(\chi^*)^i = \sigma. \quad (14)$$

Because $\sigma > 0$ and $\sum_{i=2}^{\sum_{j=1}^{M} n_j}(i-1)G_i\chi^i$ is a monotonic increasing function with values ranging from 0 to $+\infty$, when χ varies from 0 to $+\infty$, the optimal χ^* must exist and be unique. From (13), it can be seen that $(F(\chi)/G(\chi))' > 0$ when $\chi < \chi^*$ and that $(F(\chi)/G(\chi))' < 0$ when $\chi > \chi^*$. Therefore, the throughput function reaches the *maximum* value when $\tau_1^*/(1-\tau_1^*) = \chi^*$. Of course, the optimal solution varies with the variation of parameters $\alpha_i (1 \le i \le M)$. Therefore, we denote the optimal solution as $\tau_i^*(\alpha_1, \ldots, \alpha_M)$ $(1 \le i \le M)$. ■

Up to now, the answer to the question about the existence and uniqueness of the optimal operation point is quiet clear. That is, there is one and only one optimal operation point where the maximum throughput can be achieved under the constraint of achieving target bandwidth allocation ratio $\hat{\alpha}_i$.

By using (14), the optimal operation point can be obtained numerically. However, for the purpose of real implementation and obtainment of deeper insight into the system performance, it is necessary to derive more meaningful and concise approximations for the exact formulas. From (12) and (14), we have (15), shown at the bottom of the page. From (6) and (15), it can be seen that if n_i, $T_{\mathrm{Len},i} (1 \le i \le M)$ are sufficiently large, the optimal operation point $\tau_1^*(\alpha_1, \ldots, \alpha_M) \ll 1$ (it is also true for $\tau_i^*(\alpha_1, \ldots, \alpha_M)(i > 1)$). Therefore, it is reasonable to limit the discussions to the case where $\tau_i \ll 1 (i = 1, \ldots, M)$.

Next, we try to answer the question about where the optimal operation point is and what the maximum value is for the system throughput under the optimal operation point through the next theorem.

Theorem 2: Assume that $M(\ge 1)$ types of traffic coexist in the system, with $n_i (1 \le i \le M)$ numbers of type-i traffic flows. Moreover, assume that $\tau_i/(1-\tau_i) = \alpha_i \cdot (\tau_1/(1-\tau_1))(\alpha_i > 0, 1 \le i \le M, \alpha_1 \equiv 1)$. If n_i, $T_{\mathrm{Len},i} (1 \le i \le M)$ are sufficiently large, so that the optimal operation point $\tau_i^*(\alpha_1, \ldots, \alpha_M) \ll 1 (1 \le i \le M)$, then the optimal operation point can be approximated as

$$\tau_1^*(\alpha_1, \ldots, \alpha_M) \approx \frac{1}{\sqrt{\frac{T_c^*}{2}} \cdot \sum_{j=1}^{M} \alpha_j n_j} \equiv \tau_{1\_ap}^*(\alpha_1, \ldots, \alpha_M) \quad (16)$$

where $T_c^* \equiv T_c/\sigma$. Moreover, if $T_{\mathrm{Len},1} = \cdots = T_{\mathrm{Len},M} = T_{\mathrm{Len}}$ and T_{Len}, $n_i (1 \le i \le M)$ are sufficiently large,

$$\begin{aligned} S &= \frac{\sum_{i=1}^{M} n_i \cdot \frac{\tau_i}{1-\tau_i} \cdot Q\left(c_{j(1\le j\le M)} = 0\right) \cdot T_{\mathrm{Len},i}}{\left\{ Q\left(c_{j(1\le j\le M)} = 0\right)\cdot\sigma + \sum_{i=1}^{M} n_i \cdot \frac{\tau_i}{1-\tau_i} \cdot Q\left(c_{j(1\le j\le M)} = 0\right) \cdot T_{s,i} + \sum_{0\le c_j \le n_j (1\le j\le M), \sum_{j=1}^{M} c_j \ge 2} Q(c_1, \ldots, c_M) \cdot T_c(c_1, \ldots, c_M) \right\}} \\ &= \frac{F_1 \cdot \chi}{\sigma + \sum_{i=1}^{\sum_{j=1}^{M} n_j} G_i \cdot \chi^i} \\ &\equiv \frac{F(\chi)}{G(\chi)} \end{aligned} \quad (12)$$

$$\frac{\tau_1^*(\alpha_1, \ldots, \alpha_M)}{1 - \tau_1^*(\alpha_1, \ldots, \alpha_M)} \le \sqrt{\frac{\sigma}{G_2}} = \sqrt{\frac{\sigma}{\sum_{0\le c_l \le n_l (1\le l\le M),\ \sum_{l=1}^{M} c_l = 2} T_c(c_1, \ldots, c_M) \cdot \prod_{i=1}^{M} \alpha_i^{c_i} \binom{n_i}{c_i}}} \quad (15)$$

the corresponding achieved maximum throughput can be approximated as

$$S_{\max} \approx \frac{T_{\text{Len}}}{T_s + \sigma K + T_c\left[K(e^{1/K}-1)-1\right]} \tag{17}$$

where $K \equiv \sqrt{T_c^*/2}$.

Proof: According to Theorem 1, because at the optimal operation point $\tau_i^*(\alpha_1,\ldots,\alpha_M) \ll 1 (1 \le i \le M)$, it is reasonable for us to limit the discussion only within the range of $\tau_i \ll 1 (1 \le i \le M)$. Moreover, in this case, the relationship $\tau_i/(1-\tau_i) = \alpha_i \cdot (\tau_1/(1-\tau_1))$ can be further approximated as $\tau_i \approx \alpha_i \cdot \tau_1$, which is used in the following derivations.

First, we verify that T_c that is defined in (7) can be approximately regarded as a constant, if we neglect the case that three or more packets collide with each other at the same time. We have the following approximation:

$$T_c \approx \frac{\left(\begin{array}{l}\displaystyle\sum_{1\le i\le M, 1\le j\le M, i\ne j} n_i n_j \alpha_i \alpha_j \cdot T_c(c_1,\ldots,c_M) \\ \displaystyle+\sum_{i=1}^{M} n_i \cdot (n_i-1)\cdot \alpha_i^2 \cdot T_c(c_1,\ldots,c_M)\end{array}\right)}{\displaystyle\sum_{1\le i\le M, 1\le j\le M, i\ne j} n_i n_j \alpha_i \alpha_j + \sum_{i=1}^{M} n_i\cdot(n_i-1)\cdot\alpha_i^2}. \tag{18}$$

From the preceding approximation, it can be seen that once $T_{\text{Len},i}, n_i, \alpha_i (i=1,\ldots,M)$ are fixed, T_c can be approximated as a constant, which is independent of $\tau_i (1 \le i \le M)$.

Based on the assumption that $\tau_i^*(\alpha_1,\ldots,\alpha_M) \ll 1 (i = 1,\ldots,M)$ and (8), (4) can be approximated as (19), shown at the bottom of the page.

Instead of directly finding the optimal operation point by using function $S(\tau_1,\ldots,\tau_M)$ as an approximation, we determine the approximate optimal operation point by using the approximate function $f(\tau_1)/g(\tau_1)$. The optimal solution must satisfy the following condition:

$$\frac{f(\tau_1^*)}{f'(\tau_1^*)} = \frac{g(\tau_1^*)}{g'(\tau_1^*)} \tag{20}$$

since

$$\frac{dp_1}{d\tau_1} \approx (1-p_1)\cdot\sum_{i=1}^{M} n_i\alpha_i. \tag{21}$$

After substituting (21) into (20) and making some simplifications, one obtains

$$\begin{aligned}\tau_1^* T_c^* \cdot \sum_{i=1}^{M} n_i\alpha_i &= (1-p_1)|_{\tau_1=\tau_1^*}\cdot(1-T_c^*)+T_c^* \\ &\approx (1-T_c^*)\prod_{i=1}^{M}(1-\alpha_i\tau_1^*)^{n_i}+T_c^*. \end{aligned}\tag{22}$$

Because $(1-\alpha_i\tau_1^*)^{n_i} \approx 1-\alpha_i n_i \tau_1^* \approx (1-\tau_1^*)^{\alpha_i n_i}$, the preceding equation can be further approximated as

$$\tau_1^* T_c^* \cdot \sum_{i=1}^{M} n_i\alpha_i = (1-\tau_1^*)^{\sum_{i=1}^{M} n_i\alpha_i}\cdot(1-T_c^*)+T_c^*. \tag{23}$$

When there is only one type of traffic, the preceding equation is actually the same as [10, eq. (27)]. Therefore, (16) can be obtained by directly referring to [10, eq. (28)].

Next, we evaluate the maximum throughput that can be achieved. Under the condition that n_i, $T_{\text{Len},i} (1 \le i \le M)$ are sufficiently large, we substitute the approximate optimal solution $\tau_{1\_\text{ap}}^*(\alpha_1,\ldots,\alpha_M)$ into (4). Then, we have (24), shown at the bottom of the page. Because $n_i (1 \le i \le M)$

$$S \approx \frac{\tau_1 \cdot \sum_{i=1}^{M} n_i\alpha_i \cdot T_{\text{Len},i}}{\left\{\sigma + \tau_1 \cdot \sum_{i=1}^{M} n_i\cdot\alpha_i\cdot T_{s,i} + \left(\frac{1}{1-p_1} - 1 - \tau_1\cdot\sum_{i=1}^{M} n_i\alpha_i\right)\cdot T_c\right\}} \equiv \frac{f(\tau_1)}{g(\tau_1)} \tag{19}$$

$$S_{\max} \approx \frac{\tau_{1\_ap}^*\cdot\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}\cdot\sum_{i=1}^{M}\alpha_i n_i\cdot T_{\text{Len},i}}{\left\{\begin{array}{l}\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}\cdot\sigma+\tau_{1\_ap}^*\cdot\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}\cdot\sum_{i=1}^{M}\alpha_i n_i\cdot T_{s,i} \\ +\left[1-\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}-\tau_{1\_ap}^*\cdot\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}\cdot\sum_{i=1}^{M}\alpha_i n_i\right]\cdot T_c\end{array}\right\}} \tag{24}$$

are assumed to be sufficiently large, we have the following approximation:

$$\left[1-1\Big/\left(K\cdot\sum_{i=1}^{M}\alpha_i n_i\right)\right]^{\sum_{i=1}^{M}\alpha_i n_i}\approx e^{-1/K}. \quad (25)$$

Moreover, it is assumed that $T_{\mathrm{Len},1}=\cdots=T_{\mathrm{Len},M}=T_{\mathrm{Len}}$ and hence $T_{s,1}=\cdots=T_{s,M}=T_s$; then, (24) can be further approximated as (17). ■

It should be noted that (17) has no relationship with α_i, which indicates that it can be used to approximately express the global maximum value for the throughput function $S(\tau_1,\ldots,\tau_M)$ in (4) under the condition that n_i, $T_{\mathrm{Len},i}(1\le i\le M)$ are sufficiently large. Moreover, compared with [10, eq. (31)], we find that *the maximum throughput that is achieved is exactly the same no matter how many different types of traffic flows coexist in the system.*

Deduction 1: Assume that there are $M(\ge 1)$ types of traffic coexisting in the system, with $n_i(1\le i\le M)$ numbers of type-i traffic flows. Moreover, assume that $\tau_i/(1-\tau_i)=\alpha_i\cdot(\tau_1/(1-\tau_1))(\alpha_i>0, 1\le i\le M, \alpha_1\equiv 1)$. If n_i, $T_{\mathrm{Len},i}(1\le i\le M)$ are sufficiently large, so that the optimal operation point $\tau_i^*(\alpha_1,\ldots,\alpha_M)\ll 1(1\le i\le M)$, then the system operates close to the optimal operation point if and only if the packet collision rate is approximately equal to $1-e^{-1/K}(K\equiv\sqrt{T_c^*/2})$.

Proof: Assume that the system works under the optimal operation point. By substituting the optimal solution in (16) into (8), we have the following approximations:

$$\begin{aligned}(1-p_1)&\approx\cdots\\&\approx(1-p_M)\\&\approx\prod_{i=1}^{M}\left(1-\alpha_i\tau_{1\_ap}^*\right)^{n_i}\\&\approx\left(1-\tau_{1\_ap}^*\right)^{\sum_{i=1}^{M}\alpha_i n_i}\\&\approx e^{-1/K}.\end{aligned} \quad (26)$$

Therefore, the packet collision rate corresponding to the optimal operation point can be expressed as

$$p_i\approx 1-e^{-1/K},\qquad i=1,2,\ldots,M. \quad (27)$$

Because the function in (8) is continuous and monotonic, it is easy to arrive at the conclusion that if the packet collision rate is approximately equal to $1-e^{-1/K}$, the system must operate close to the optimal operation point. ■

Again, this conclusion is general because it does not depend on parameters $\alpha_i(1\le i\le M)$. We can use the preceding equation to check if the system works under or close to the optimal operation point.

TABLE I
SYSTEM PARAMETERS

Channel Bit Rate	272 bits
PHY Header	192 μs
ACK	112 bits +PHY header
Channel Bit Rate	11Mbps
Propagation Delay	1 μs
Slot Time	20 μs
SIFS	10 μs
DIFS	50 μs

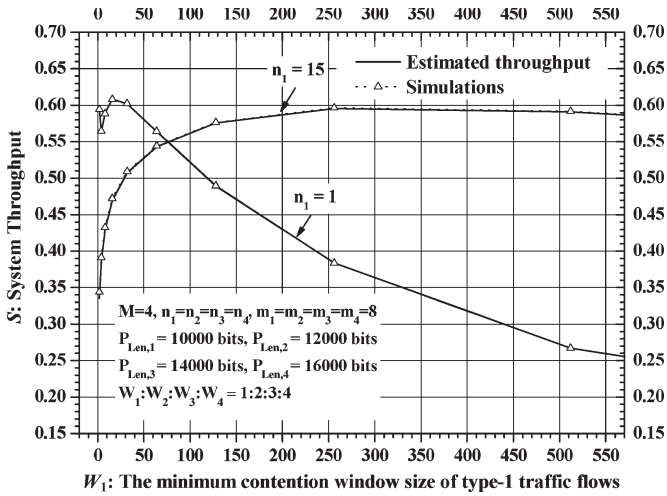


Fig. 2. Verification of fundamental equations (1)–(4).

V. RESULT AND DISCUSSION

In the previous section, we analyze the optimal operation point and its corresponding maximum throughput. In this section, we verify some important approximated results that are obtained by using both simulation and numerical methods. In our examples, the parameters for the system are summarized in Table I based on IEEE 802.11b.

Since the maximum throughput analysis is based on fundamental equations (1)–(4), in the first example, we verify these equations by using simulations. In this paper, all the discrete event simulations are developed and executed over an OPNET Modeler. In this example, a single-hop system is considered, where there are $M=4$ types of traffic, with $n_1=n_2=n_3=n_4$. The packet payload sizes for different traffic flows are assigned as $P_{\mathrm{Len},1}=10\,000$ bit, $P_{\mathrm{Len},2}=12\,000$ bit, $P_{\mathrm{Len},3}=14\,000$ bit, and $P_{\mathrm{Len},4}=16\,000$ bit. The ratios among minimum CW sizes for different traffic flows are set as $W_1:W_2:W_3:W_4=1:2:3:4$. Moreover, it is assumed that the channel conditions are ideal (i.e., no hidden terminals and capture). In simulations, system throughput S is obtained by varying the minimum CW sizes $W_i(i=1,2,3,4)$. On the other hand, in order to verify (1)–(4), throughput S is numerically calculated using the following procedure: First, packet sending rates $\tau_i(i=1,2,3,4)$ and packet collision rates $p_i(i=1,2,3,4)$ are calculated based on (1) and (2). Then, the throughput is estimated by substituting $\tau_i(i=1,2,3,4)$ and $p_i(i=1,2,3,4)$ into (4). The throughput obtained by using simulation and numerical ways are compared in Fig. 2. In the

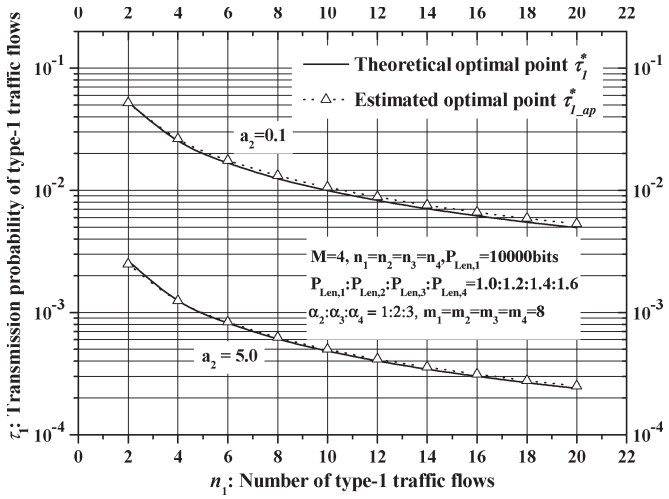


Fig. 3. Comparisons between the theoretical optimal operation points and the estimated ones.

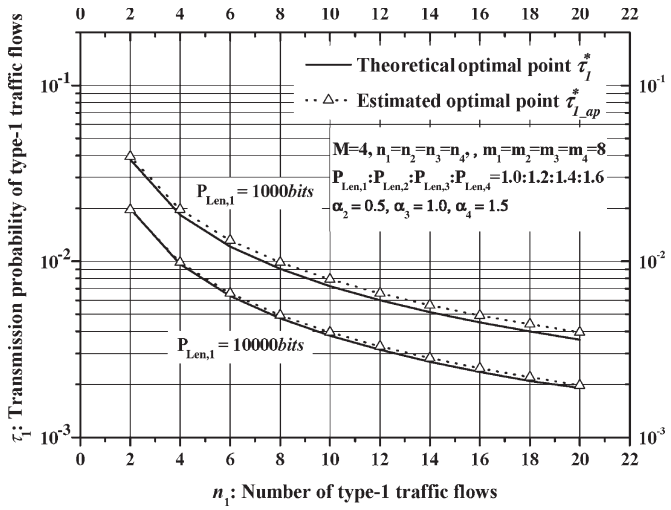


Fig. 4. Comparisons between the theoretical optimal operation points and the estimated ones.

figure, two cases are shown: One is $n_1 = 1$, and the other is $n_1 = 15$. Comparisons show that the numerical results agree with the simulation results well, which verifies the fundamental mathematical relationships that are given in (1)–(4). In Fig. 2, for the case of $n_1 = 1$, it seems that there are more than one local optimal point for the throughput. One is near $W_1 = 2$, and the other one is near $W_1 = 16$. In Theorem 1, we point out that there is only one optimal point, which corresponds to the maximum throughput. Do the results in the figure suggest a contradiction with Theorem 1? The answer is no. This is because, in Theorem 1, we give a very important premise, i.e., parameters $\alpha_i (1 \leq i \leq M)$ should be given and kept as constants. However, in the simulation that is shown in Fig. 2, with the increase in CW sizes, the corresponding parameters $\alpha_i (1 \leq i \leq M)$ vary. Therefore, in this case, the premise for Theorem 1 does not hold.

In the second example, we compare the exact optimal operation points τ_1^* that are numerically obtained from (4) with the approximated optimal operation points $\tau_{1\_ap}^*$ that are obtained from (16). In the example, four types of traffic are considered, with $\alpha_2 : \alpha_3 : \alpha_4 = 1 : 2 : 3$. In Fig. 3, comparison results of optimal operation points are shown versus the number of type-1 traffic flows n_1. Two cases are shown: One is $\alpha_2 = 0.1$, and the other one is $\alpha_2 = 5.0$. From the figure, it can be seen that good agreements between exact and approximate optimal operation points can be achieved if the number of traffic flows n_1 is not so small. Furthermore, comparisons between the cases of $\alpha_2 = 0.1$ and $\alpha_2 = 5.0$ show that good estimation accuracy can be obtained as long as the estimated optimal operation point is far less than one, which is the condition we based on when proving the approximate estimation in (16).

Moreover, in Fig. 4, we make further comparisons between the exact optimal operation points and approximated ones under the cases that $P_{\mathrm{Len},1} = 1000$ bit and $P_{\mathrm{Len},1} = 10\,000$ bit. Other parameters are shown in the figure. Again, it can be seen that good agreements between exact and approximate optimal operation points can be achieved as long as the estimated optimal operation point is far less than one.

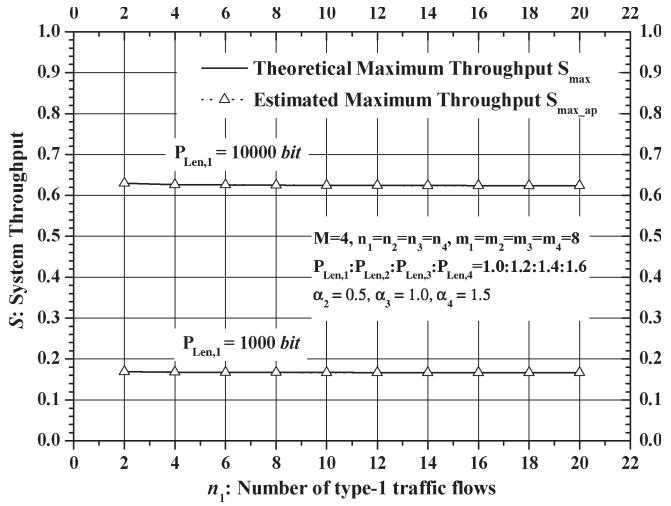


Fig. 5. Comparisons between the theoretical maximum throughput and the estimated ones.

After verifying the accuracy of the estimation of the optimal operation point, we illustrate the accuracy of the evaluated maximum throughput by using the estimated optimal operation point. In order to obtain the exact maximum throughput and its evaluated value, we substitute the exact optimal operational point and its corresponding approximated one into (4). The comparison results are given in Fig. 5 under the cases that $P_{\mathrm{Len},1} = 1000$ bit and $P_{\mathrm{Len},1} = 10\,000$ bit. It can be seen that the estimated maximum throughput $S_{\max\_ap}$ agrees with the corresponding theoretical value $S_{\max}$ well. One reason for this is that the accuracy of the estimated optimal operation point is high. The other reason is that function $S(\tau_1, \ldots, \tau_M)$ is smooth; even a nonnegligible difference in the estimation of the optimal operation point leads to similar throughput values.

In (17), a simple approximated estimation on the maximum throughput is given for the case where the packet payload lengths for all types of traffic flows are equal, i.e., $P_{\mathrm{Len},1} = P_{\mathrm{Len},2} = \cdots = P_{\mathrm{Len},M} (M \geq 1)$. Comparison results between

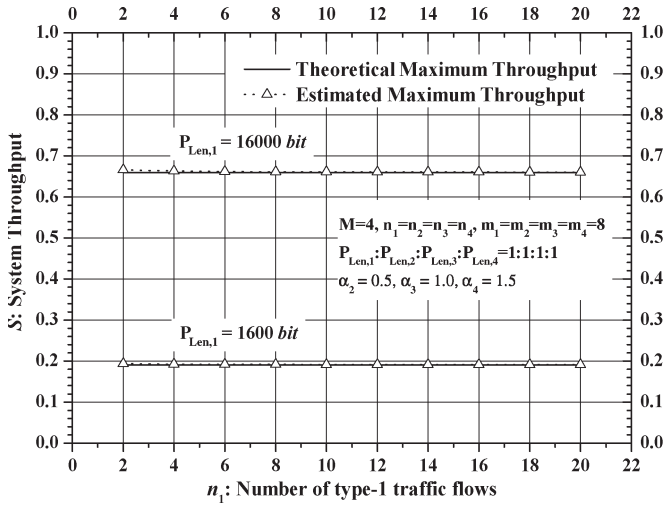


Fig. 6. Comparisons between the theoretical maximum throughput and the estimated ones for the case where the packet payload lengths for all types of traffic flows are equal.

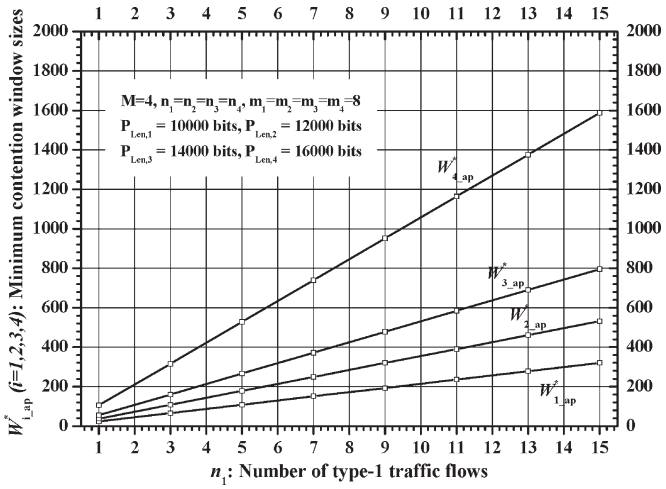


Fig. 7. Numerical results for the optimal minimum CW sizes $W^*_{i\_ap}(1 \le i \le M)$.

the theoretical maximum throughput and the approximated estimated maximum throughput, which is obtained from (17), are shown in Fig. 6. It can be seen that the estimated maximum throughput agrees with the corresponding theoretical value well under the case where the number of traffic flows is not so small.

VI. Achieving the Maximum Throughput and Service Differentiation: Basic Idea

In the previous section, the basic theoretical results that are proposed in this paper are verified by using both simulation and numerical ways. In this section, basic ideas for achieving maximum system throughput S and target service differentiations among different types of traffic flows, i.e., $\hat{\alpha}_i \equiv s_i/s_1(\hat{\alpha}_i > 0, 1 \le i \le M)$, are described and verified by using both simulation and numerical ways.

Given target bandwidth differentiations $\hat{\alpha}_i(1 \le i \le M)$, based on (11), the ratios for the packet sending rates between different traffic flows $\alpha_i \equiv ((\tau_i/(1-\tau_i))/(\tau_1/(1-\tau_1)))(1 \le i \le M)$ can be given as

$$\alpha_i = \hat{\alpha}_i \cdot T_{\mathrm{Len},1}/T_{\mathrm{Len},i}. \tag{28}$$

Then, the approximated optimal operation point $\tau^*_{i\_ap}(\alpha_1, \ldots, \alpha_M)(1 \le i \le M)$, where the maximum throughput and target bandwidth differentiations can be achieved, can be obtained by combining (16) in Theorem 2 with (28), i.e.,

$$\begin{cases} \tau^*_{1\_ap}(\alpha_1, \ldots, \alpha_M) = \dfrac{1}{\sqrt{\frac{T^*_c}{2} \cdot \sum_{j=1}^{M} \alpha_j n_j}} \equiv \dfrac{1}{\sqrt{\frac{T^*_c}{2} \cdot E_1}} \\ \tau^*_{i\_ap}(\alpha_1, \ldots, \alpha_M) = \dfrac{\alpha_i \left(\frac{\tau^*_{1\_ap}}{1-\tau^*_{1\_ap}} \right)}{1+\alpha_i \left(\frac{\tau^*_{1\_ap}}{1-\tau^*_{1\_ap}} \right)}, \qquad i = 1, \ldots, M \end{cases} \tag{29}$$

where $E_1(\equiv \sum_{j=1}^{M} \alpha_j n_j)$. Based on (8), the corresponding packet collision rates $p^*_{i\_ap}(1 \le i \le M)$ can be given as

$$p^*_{i\_ap} = 1 - \left(1 - \tau^*_{i\_ap}\right)^{n_i - 1} \prod_{j=1, j \ne i}^{M} \left(1 - \tau^*_{j\_ap}\right)^{n_j}. \tag{30}$$

Finally, based on (1) and (2), minimum CW sizes $W^*_{i\_ap}(1 \le i \le M)$ corresponding to the optimal operation point can be set as follows:

$$W^*_{i\_ap} \approx \frac{2\left(1 - 2p^*_{i\_ap}\right)}{\left(1 - 2p^*_{i\_ap}\right)\tau^*_{i\_ap} + p^*_{i\_ap} \cdot \tau^*_{i\_ap}\left[1 - \left(2p^*_{i\_ap}\right)^{m_i}\right]}. \tag{31}$$

Next, simulation results are given to verify (28)–(31). In order to verify the accuracy of the relationships that are proposed in the preceding equations, the corresponding $W^*_{i\_ap}(1 \le i \le M)$ are calculated given the target service-differentiation ratios $\hat{\alpha}_i(1 \le i \le M)$ by using (28)–(31). Then, in the simulations, minimum CW sizes $W_i(1 \le i \le M)$ are set to be equal to the numerically obtained $W^*_{i\_ap}(1 \le i \le M)$. Finally, the achieved throughput and bandwidth ratios between different traffic flows that are measured from the simulations are compared with the corresponding theoretical maximum throughput $S_{\max}$ and target bandwidth differentiations $\hat{\alpha}_i(1 \le i \le M)$. In the simulations, four different types of traffic flows are considered. The target bandwidth differentiation ratios are set as $\hat{\alpha}_2 = 0.75$, $\hat{\alpha}_3 = 0.56$, and $\hat{\alpha}_4 = 0.32$.

In Fig. 7, by using (28)–(31), the obtained numerical results for minimum CW sizes $W^*_{i\_ap}(1 \le i \le M)$ through which the maximum throughput and target service differentiations can be achieved are shown. Fig. 8 shows comparisons between the theoretical maximum throughput and the achieved throughput that is measured from simulation by setting $W_i = W^*_{i\_ap}(1 \le i \le M)$. Fig. 9 shows comparisons between the target bandwidth differentiations $\hat{\alpha}_i(\hat{\alpha}_i > 0, 1 \le i \le M)$ and the achieved

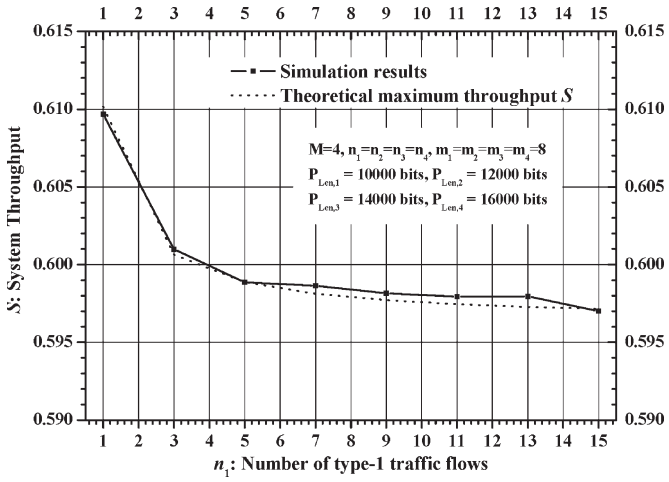


Fig. 8. Comparisons between the theoretical maximum throughput and the achieved throughput.

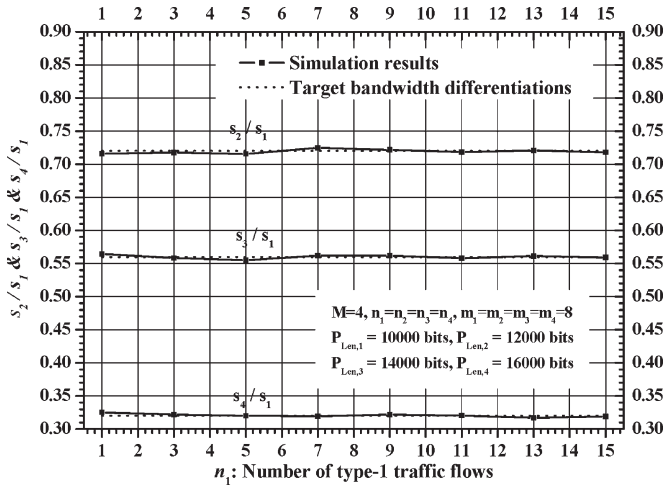


Fig. 9. Comparisons between the target bandwidth differentiations and the achieved ones.

bandwidth ratios between different traffic flows. Based on these comparisons, it can be concluded that by using the relationships that are given in (28)–(31), the maximum throughput and target bandwidth differentiations can be achieved at the same time.

VII. Achieving the Maximum Throughput and Service Differentiation: Adaptive Scheme

In this section, based on the preceding analysis, we propose a simple adaptive scheme to achieve the maximum throughput and maintain target bandwidth differentiation between different types of traffic flows. It is assumed that the system is at saturation state with ideal channel conditions. In IEEE 802.11 [4], it is specified that data packets that are generated by a higher protocol layer are fragmented into smaller MAC layer frames for transmission. Therefore, it is reasonable to assume that each traffic flow has the same packet payload length P_{Len} for the MAC frame. Moreover, with all the traffic flows adopting the same MAC packet payload length, some theoretical results that are proposed in this paper can be reduced into

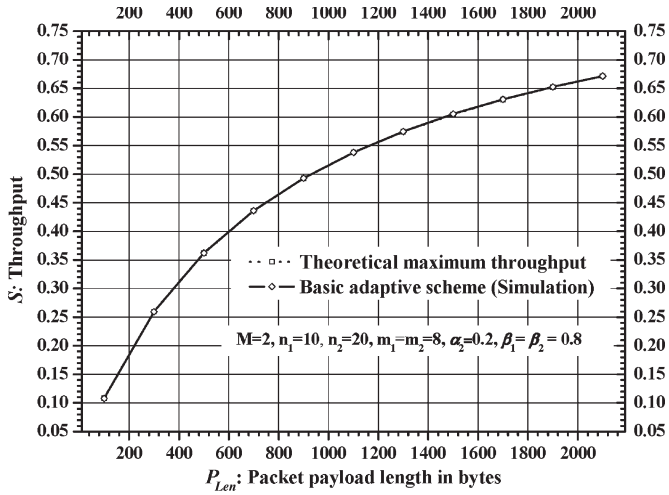


Fig. 10. Comparisons between the theoretical maximum throughput and the simulated ones, which are archived by using the basic adaptive scheme.

simpler forms, which helps to simplify their implementations in a real-world system. However, by following the same way that is given in this section, the proposed adaptive scheme can be easily extended to more general cases based on the proposed general theoretical results in this paper.

A. Basic Adaptive Scheme

In the basic adaptive scheme, it is assumed that $n_i (1 \leq i \leq M)$ and $\hat{\alpha}_i (1 \leq i \leq M)$ are known by each sending station in advance. Moreover, for simplicity, all the traffic flows adopt the same MAC packet payload length.

According to deduction 1, packet collision rates at the optimal operation point $p^*_{i\_ap} (1 \leq i \leq M)$ can be simply evaluated as

$$p^*_{i\_ap} \approx 1 - e^{-\sqrt{2\sigma/T_c}}. \tag{32}$$

In order to achieve the maximum throughput and the target bandwidth differentiation, optimal minimum CW sizes $W^*_{i\_ap} (1 \leq i \leq M)$ can be obtained by combining (29), (31), and (32). Finally, $W^*_{i\_ap} (1 \leq i \leq M)$ are used to adjust the current minimum CW sizes Current_$W_i (1 \leq i \leq M)$ as follows:

$$\text{Current\_}W_i = \beta_i \cdot \text{Current\_}W_i + (1 - \beta_i) \cdot W^*_{i\_ap} \tag{33}$$

where $1 \leq i \leq M$. $\beta_i \in [0, 1]$ is a smoothing factor.

Simulations have been done to verify the performance of the preceding scheme. Without loss of generality, two different traffic types are considered. The number of traffic flows is set as $n_1 = 10$ and $n_2 = 20$, which is known by each sending station in the current BSS. No central coordinator (CC) is needed. System parameters are set according to Table I. Specifically, $\alpha_2 = 0.2$, and $\beta_1 = \beta_2 = 0.8$. Both type-1 and type-2 traffic flows begin their minimum CW size from 512.

Fig. 10 shows the comparison between the theoretical maximum throughput $S_{\max}$ and the actual throughput S that is

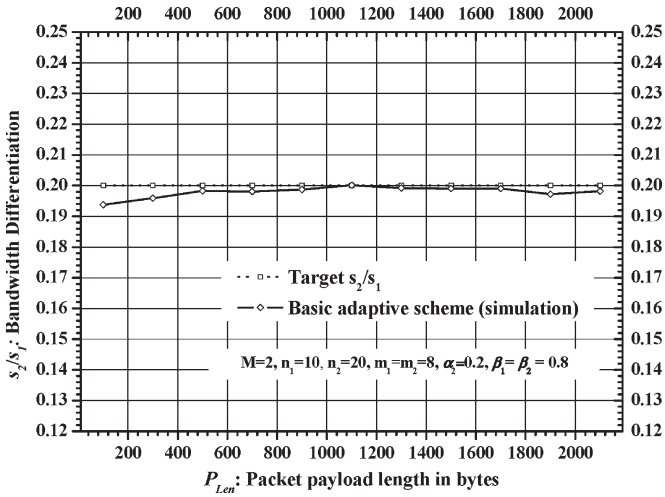


Fig. 11. Bandwidth differentiations that are achieved by the basic adaptive scheme.

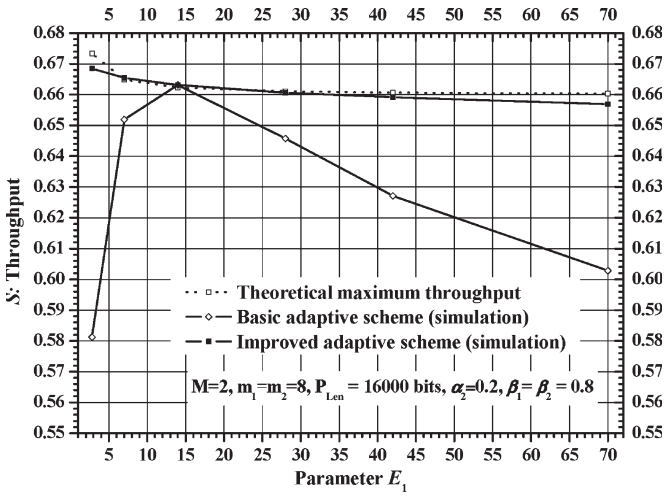


Fig. 12. Throughput that are achieved by the basic adaptive scheme and its improved version.

achieved by using the basic adaptive scheme. Moreover, the achieved s_2/s_1 is shown in Fig. 11. It can be seen that the proposed adaptive scheme achieves the maximum throughput and, at the same time, the target service-differentiation performance.

In the basic adaptive scheme, it is assumed that $n_i (1 \leq i \leq M)$ are known by each station. According to (29), it is evident that if the exact value of $n_i (1 \leq i \leq M)$ is different from the assumed ones, the achieved throughput will deviate from the maximum throughput. However, extensive experiments show that the sensitivity of the achieved throughput to change in the number of traffic flows is not so high. In Fig. 12, the sensitivities of the achieved throughput to the variations of parameter E_1 are shown (see the "Basic adaptive scheme" in the figure). In the experiments, the mobile stations use the fixed setting of $E_1 = 14.0$ and $\alpha_2 = 0.2$ to calculate the target optimal operation points $\tau^*_{1\_ap}$ and $\tau^*_{2\_ap}$. We change the number of mobile stations n_1 and n_2; hence, the actual values of E_1 change from 2.8 to 70.0. From the figure, it can be seen that the achieved throughput does not deviate much from

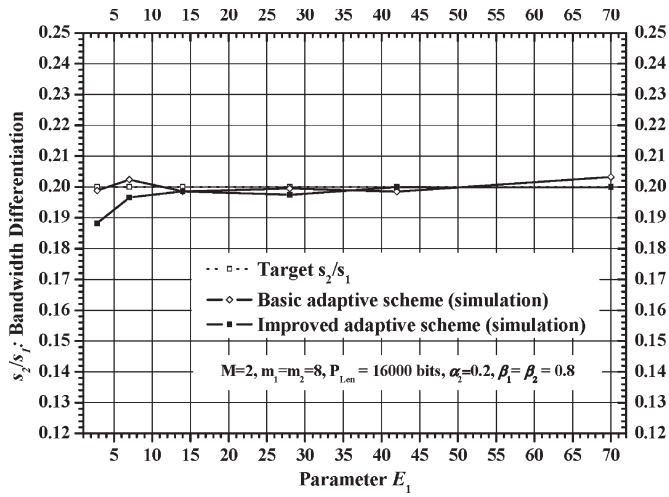


Fig. 13. Bandwidth differentiations that are achieved by the basic adaptive scheme and its improved version.

the corresponding maximum throughput if the actual value of E_1 does not deviate much from the assumed value of 14.0. Therefore, it can be found that the achieved throughputs are not very sensitive to the variation of the number of traffic flows, i.e., for example, to some extent, the system can achieve optimal performance by using the basic adaptive scheme; even the actual number of traffic flows is different from the assumed one. This is because the throughput function that is defined in (4) is relatively smooth versus the variation of packet sending rates τ_i. However, it can be also seen that if the actual value of E_1 deviates much from the assumed one, the achieved throughput is far less than the ideal maximum throughput, which indicates lower utilization of the system (see the cases for $E_1 = 2.8$ or $E_1 = 70.0$). Moreover, from Fig. 13, it can be seen that, in the basic adaptive scheme, service differentiation s_2/s_1 is not influenced much by the variations of E_1.

B. Improved Adaptive Scheme

A centralized version of the adaptive scheme using a CC is proposed in this section. In our scheme, the CC itself carries traffic flows for transmission (which we assume are of type-1), and in addition, it serves as a coordinator to guarantee that the centralized knowledge can be used to achieve the maximum throughput and target service differentiation even in a dynamic context, when the number of active mobile stations changes.

The functions of a CC in the improved scheme can be explained as follows: It detects the value of E_1 [see (29)] at run time. If the detected value of E_1 is somewhat far from the assumed ones, the CC broadcasts new values, which are the corresponding averaged values of the newly detected E_1. In order to maintain the target bandwidth differentiation between different traffic flows, it is important for the CC to broadcast E_1, together with the target bandwidth differentiation ratios $\hat{\alpha}_i (1 \leq i \leq M)$. Receiving these updated values, all the stations in the current BSS update their memorized values of E_1 and $\hat{\alpha}_i (1 \leq i \leq M)$. Finally, $W^*_{i\_ap} (1 \leq i \leq M)$, which can be obtained by using (29), (31), and (32) one by one, are used

to adjust the current minimum CW sizes Current_$W_i(1 \le i \le M)$ [see (33)]. It can be seen that, for normal stations, their adaptive schemes are almost unchanged compared with the one in the basic adaptive scheme. They only need to change their memorized E_1 and $\hat{\alpha}_i(1 \le i \le M)$ if new broadcasted values are obtained.

In order to keep track of the change in the number of active mobile stations, the CC monitors the traffic on the channels and evaluates the real-time value for E_1. In the following, we propose a method to evaluate the real-time value for E_1. In the case where $\tau_i \ll 1(1 \le i \le M)$ and $\tau_i \approx \alpha_i \tau_1 (2 \le i \le M)$, and combining with (8), we have

$$(1-p_1) \approx (1-\tau_1)^{\sum_{j=1}^{M} \alpha_j n_j} = (1-\tau_1)^{E_1}. \tag{34}$$

Therefore, the estimation of E_1 can be given as

$$\hat{E}_1 = \log(1-p_1)/\log(1-\tau_1) \tag{35}$$

where packet collision rate p_1 can be easily evaluated at run time. In [30], an efficient way to evaluate the run-time packet collision rate is proposed. τ_1 can be obtained by substituting the estimated p_1 and the current minimum CW size Current_W_1 into (1). After obtaining $\hat{E}_1$, it should be averaged as $\bar{E}_1$ and compared with Current_E_1, which is the current memorized value for E_1. $\bar{E}_1$ can be expressed as

$$\bar{E}_1 = \beta_1 \cdot \bar{E}_1 + (1-\beta_1) \cdot \hat{E}_1. \tag{36}$$

If $\bar{E}_1$ is less than Current_$E_1 \cdot \gamma (0 < \gamma < 1)$ during the past $k_t \ge 1$ comparisons, Current_E_1 is set as $\bar{E}_1$. If $\bar{E}_1$ is larger than Current_$E_1/\gamma (0 < \gamma < 1)$ during the past $k_t \ge 1$ comparisons, Current_E_1 is updated as $\bar{E}_1$.

In the scheme, if γ is set to be very near zero, the improved scheme is actually the same as the basic scheme. On the other hand, if γ is set very near one, the CC will modify $E_i(1 \le i \le M)$ too often, which proves to be unnecessary according to the former discussions on the sensitivities of the achieved throughput to the variation of the number of traffic flows. Therefore, parameter γ should be carefully chosen to improve the performance of the system and, at the same time, to minimize the control overhead. For the same reason, parameter k_t must also be carefully chosen to avoid heavy increase in the control overhead.

The performance of the improved scheme is verified by simulation. In the simulation, a station carrying type-1 traffic flow is chosen to serve as the CC. γ and k_t are set to be 0.5 and 10, respectively. If the CC decides to broadcast new $E_i(1 \le i \le M)$, it generates a special management frame and gains access to the channel by using the highest medium access priority (PIFS) to ensure that the new values can be received by other traffic flows as soon as possible.

In Figs. 12 and 13, performance comparisons between the basic adaptive scheme and the improved one are shown. From Fig. 12, it can be seen that, by using the improved scheme, the achieved throughput S is closer to the corresponding maximum throughput $S_{\max}$ for all cases, which is caused by the ability to dynamically adapt to changing values of $E_i(1 \le i \le M)$. Moreover, from Fig. 13, it can be seen that service differentiation, which is measured by s_2/s_1, is kept. For the case of a small E_1, for example, one that is smaller than 5.0, it can be seen that the s_2/s_1 in the improved adaptive scheme further deviates from the target value of 0.2. This is because the estimation of E_1 is based on the premise that packet sending rates $\tau_i \ll 1(i = 1, 2)$. However, in the case of a small E_1, $\tau_i(i = 1, 2)$ are larger than those in the basic adaptive scheme; thus, the estimation errors for E_1 cannot be neglected. On the other hand, for the case of larger E_1, $\tau_i(i = 1, 2)$ are far less than 1.0, and the estimation of E_1 is more accurate, which makes the achieved service differentiation near the target value. Therefore, in this case, by using the improved adaptive scheme, system performance, which is measured in both the throughput and service differentiation, almost reaches their optimal values.

VIII. Conclusion

We propose an analysis model for computing the system throughput. Moreover, based on the model, we derive approximations to get simpler but more meaningful relationships among different parameters. The significant contribution of this paper is that we successfully analyze the optimal operation point where the maximum throughput can be achieved. Moreover, we propose a simple adaptive scheme that can make the system operate under the optimal operation point and achieve target service differentiation between different traffic flows.

In our future work, we would like to consider how to extend the results that are obtained in this paper to other research topics.

1) The nonsaturation state should be considered. In this paper, performance analysis is based on the assumption that the system is at a saturation state. A real-world system mostly works at a nonsaturation state. In our recent work [31], it is found that Bianchi's model [10] can be extended to describe the system performance characteristics in a nonsaturation state, which suggests that it is possible to combine researches on saturation state and nonsaturation state together.
2) Nonideal channel conditions should be considered in future work. In order to optimize the system performance in more practical channel conditions, the performance analysis model that is proposed in this paper should be extended to consider some more practical situations, such as transmission errors, hidden terminals, and captures.
3) Performance characteristics of different service differentiation supporting mechanisms should be studied and compared further. In this paper, only one mechanism for supporting service differentiation is studied, i.e., differentiating the minimum CW sizes according to the priority of different traffic categories. However, in available literatures such as the IEEE 802.11e EDCA, more than one service-differentiation-supporting mechanism is proposed. Therefore, the advantages and disadvantages of these mechanisms and how these mechanisms can work together are important and interesting research work.

4) Performance optimization in systems where both DCF and PCF are supported is another interesting research topic. In this paper, only DCF is considered. However, PCF has its own advantages in supporting real-time traffic. On the other hand, DCF serves as the basis for the IEEE 802.11 MAC protocol and has already been widely accepted by product vendors. Moreover, enhanced DCF is also equipped with its own QoS supporting mechanisms, such as EDCA in IEEE 802.11e. Therefore, to our opinion, we believe that future performance optimization should consider both DCF and PCF, which requires more sophisticated research.

We believe that the results that are proposed in this paper will serve as a solid basis for future possible extensions.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments to improve the quality of this paper.

REFERENCES

[1] Y. Cheng and W. H. Zhuang, "DiffServ resource allocation for fast handoff in wireless mobile Internet," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 130–136, May 2002.

[2] R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet architecture: An Overview*, Jun. 1994. IETF RFC 1633.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, *An Architecture for Differential Services*, Dec. 1998. IETF RFC 2475.

[4] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11, Aug. 1999.

[5] J. Weinmiller, M. Schlager, A. Festag, and A. Wolisz, "Performance study of access control in wireless LANs IEEE 802.11 DFWMAC and ETSI RES 10 HIPERLAN," *Mobile Netw. Appl.*, vol. 2, no. 1, pp. 55–67, Jun. 1997.

[6] H. S. Chhaya and S. Gupta, "Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol," *Wirel. Netw.*, vol. 3, no. 3, pp. 217–234, 1997.

[7] T. S. Ho and K. C. Chen, "Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LAN's," in *Proc. IEEE PIMRC*, Taipei, Taiwan, R.O.C., Oct. 1996, pp. 392–396.

[8] 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, vol. 1, pp. 142–149.

[9] G. Bianchi, L. Fratta, and M. Oliveri, "Performance analysis of IEEE 802.11 CSMA/CA medium access control protocol," in *Proc. IEEE PIMRC*, Taipei, Taiwan, R.O.C., Oct. 1996, pp. 407–411.

[10] 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.

[11] Y. C. Tay and K. C. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wirel. Netw.*, vol. 7, no. 2, pp. 159–171, Mar./Apr. 2001.

[12] J. L. Sobrinho and A. S. Krishnakumar, "Distributed multiple access procedures to provide voice communications over IEEE 802.11 wireless networks," in *Proc. GLOBECOM*, 1996, pp. 1689–1694.

[13] J. Deng and R. S. Chang, "A priority scheme for IEEE 802.11 DCF access method," *IEICE Trans. Commun.*, vol. 82-B, no. 1, pp. 96–102, Jan. 1999.

[14] A. Veres, A. T. Campbell, M. Barry, and L. H. Sun, "Supporting service differentiation in wireless packet networks using distributed control," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2081–2093, Oct. 2001.

[15] I. Aad and C. Castelluccia, "Differentiation mechanisms for IEEE 802.11," in *Proc. IEEE INFOCOM*, 2001, pp. 209–218.

[16] W. Haitao, C. Shiduan, P. Yong, L. Keping, and M. Jian, "IEEE 802.11 distributed coordination function (DCF): Analysis and enhancement," in *Proc. ICC*, 2002, pp. 605–609.

[17] S. Mangold, S. Choi, P. May, O. Klein, G. Hietz, and L. Stibor, "IEEE 802.11e wireless LAN for quality of service," in *Proc. Eur. Wireless*, Feb. 2002, pp. 32–39.

[18] IEEE 802.11 WG, IEEE 802.11e/D12.0, *Draft Supplement to Standard for Telecommunications and Information Exchange Between Systems—LAN/MAN Specific Requirements—Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)*, Nov. 2004.

[19] Z. Jun, G. Zihua, Z. Qian, and Z. Wenwu, "Performance study of MAC for service differentiation in IEEE 802.11," in *Proc. IEEE GLOBECOM*, Nov. 17–21, 2002, vol. 1, pp. 778–782.

[20] B. Li and R. Battiti, "Performance analysis of an enhanced IEEE 802.11 distributed coordination function supporting service differentiation," in *Proc. Int. Workshop QoFIS*, Stockholm, Sweden, 2003, vol. LNCS 2811, pp. 152–161.

[21] B. Li and R. Battiti, "Supporting service differentiation with enhancements of the IEEE 802.11 MAC protocol: Models and analysis," Dept. Comput. Sci. Telecommun. Univ. Trento, Trento, Italy, Tech. Rep. DIT-03-024. [Online]. Available: http://eprints.biblio.unitn.it/archive/00000418/

[22] Y. Xiao, "A simple and effective priority scheme for IEEE 802.11," *IEEE Commun. Lett.*, vol. 7, no. 2, pp. 70–72, Feb. 2003.

[23] Y. Xiao, "Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1506–1515, Jul. 2005.

[24] 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.

[25] L. Romdhani, Q. Ni, and T. Turletti, "AEDCF: Enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks," INRIA Tech. Rep. 4544. [Online]. Available: http://www.inria.fr/rrrt/rr-4544.html

[26] H. Zhu, G. Cao, A. Yener, and A. D. Mathias, "EDCF-DM: A novel enhanced distributed coordination function for wireless ad hoc networks," in *Proc. IEEE ICC*, Jun. 2004, vol. 7, pp. 3886–3890.

[27] M. Benveniste, G. Chesson, M. Hoehen, A. Singla, H. Teunissen, and M. Wentink, "EDCF proposed draft text," IEEE working document 802.11-01/131r1, Mar. 2001.

[28] A. Lindgren, A. Almquist, and O. Schelen, "Evaluation of quality of service schemes for IEEE 802.11 wireless LANs," in *Proc. IEEE Conf. LCN*, Nov. 15–16, 2001, pp. 348–351.

[29] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications*. New York: Wiley-Interscience, 1998, pp. 140–144.

[30] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in *Proc. IEEE INFOCOM*, 2003, vol. 2, pp. 844–852.

[31] B. Li and R. Battiti, "Analysis of the IEEE 802.11 DCF with service differentiation support in non-saturation conditions," in *Proc. Int. Workshop QoFIS*, Barcelona, Spain, 2004, vol. LNCS 3266, pp. 64–73.

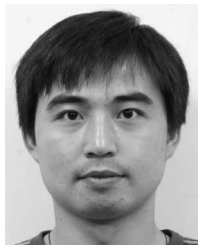
Bo Li received the B.S., M.S., and Ph.D. degrees in communications engineering from Xidian University, Xi'an, China, in 1994, 1996, and 2002, respectively.

From 2002 to 2004, he was a Postdoctoral Researcher at the University of Trento, Trento, Italy. He is currently with the School of Electronics and Information Engineering, Northwestern Polytechnic University, Xi'an, China, as a Full-Time Professor. His current research interests include multimedia wireless communication networks, cross-layer design of wireless communications systems, and resource allocations.

Roberto Battiti (M'95–A'01) received the Laurea degree from the University of Trento, Trento, Italy, in 1985, and the Ph.D. degree from the California Institute of Technology, Pasadena, in 1990.

He has been a Consultant in the area of parallel computing and pattern recognition. Since 1991, he has been a Faculty Member at the University of Trento, where he is currently a Full Professor of computer networks in the Department of Computer Science and Telecommunications. He is the author of more than 50 scientific publications, including three special issues dedicated to neural computation and experimental algorithmics. His research interests include heuristic algorithms for optimization problems, particularly, reactive search algorithms for maximum clique, maximum satisfiability, graph coloring, networks and massively parallel architectures, code assignment in wireless and cellular networks, protocols for pricing, and quality of service in wireless networks.

Prof. Battiti is a member of the Association for Computing Machinery (ACM), IEEE Computer Society, and ACM Sigmobile. He is an Associate Editor of various scientific journals and acted as Chair in the Program Committee of several international conferences. He is currently the Deputy Dean at the Faculty of Science, a Representative of the University of Trento in the Management Council of the Italian National Consortium for Computer Science (CINI), a member of the Advisory Committee for the future Telecommunications Plan of the Autonomous Province of Trento, a member of the Faculty Commission for Quality Control, and an Evaluator of the Province of Trento for projects related to computer networks and telecommunications.



Yong Fang received the B.S., M.S., and Ph.D. degree in communications engineering from Xidian University, Xi'an, China, in 2000, 2003 and 2005, respectively.

He is currently a Postdoctoral Researcher in the School of Electronics and Information, Northwestern Polytechnic University, Xi'an, China. His research interests include performance analysis of wireless networks and image/video compression and transmission.