Achieving Maximum Throughput and Service Differentiation by Enhancing the IEEE 802.11 MAC Protocol*

Bo Li and Roberto Battiti

Department of Computer Science and Telecommunications, University of Trento,
38050 POVO Trento, Italy
{li, battiti}@dit.unitn.it

Abstract. To satisfy various needs and priorities of different users and applications, Wireless LANs are currently evolving to support service differentiation. Work is in progress to define a standard enhanced version of the IEEE 802.11 Distributed Coordination Function (DCF), capable of supporting QoS for multimedia traffic at the MAC layer. This paper focuses onto one of the building blocks of this enhancement, i.e., differentiating the minimum contention window size according to the priority of different traffic categories. The novel contribution is the analysis of the optimal operation point where the maximum throughput can be achieved. The second contribution is the proposal of simple adaptive schemes which can lead the system to operate under the optimal operation point and, at the same time, achieve the target service differentiation between different traffic flows. Results obtained in the paper are relevant for both theoretical research and implementations of real systems.

1 Introduction

To provide seamless multimedia services to nomadic users and to use the spectrum in an efficient way, the “wireless mobile Internet” based on the 802.11 protocol has to provide suitable levels of Quality of Service [1]–[3]. The starting point of the paper is the IEEE 802.11 Distributed Coordination Function (DCF) standard [4], which is compatible with the current best-effort service model of the Internet, see [5]-[11] for seminal works on related models and simulations.

In order to support different QoS requirements for various types of service, a possibility is to support differentiation at the IEEE 802.11 MAC layer, as proposed in [12]-[15]. In these papers, service differentiation is achieved by assigning different minimum contention windows, different inter-frame spacing, or different maximum frame lengths to different types of traffic flows. In [16], both the Enhanced Distributed Coordination Function (EDCF) and the Hybrid Coordination Function (HCF), defined in the IEEE 802.11e draft, are extensively evaluated through simulation. In [17], the performance of the IEEE 802.11 MAC protocol with service differentiation is analyzed. However, the model is complex, which makes it difficult

* This work is supported by the project of WILMA funded by Provincia Autonoma di Trento (www.wilmaproject.org)

R. Battiti, M. Conti, and R. Lo Cigno (Eds.): WONS 2004, LNCS 2928, pp. 285–300, 2004.

to obtain deeper insight into the system performance. In [18], we propose a simple analysis model to compute the throughput in a WLAN with Enhanced IEEE 802.11 DCF.

Some more practical adaptive schemes are proposed to make the system cope with the dynamic traffic. In [19], a scheme to dynamically tune the IEEE 802.11 protocol parameters has been proposed to achieve maximum throughput. However, multiple service types are not considered. In [20], an adaptive EDCF scheme is proposed. The method uses the idea of slowly decreasing the contention window size to improve the system utilization. Service differentiation is also considered but without a rigorous analysis model to achieve maximum throughput and target service differentiation at the same time. The problem of fairly sharing channel resources is considered for example in [21]-[22] for the case of non-fully connected or ad-hoc networks. Achieving efficient utilization and weighted fairness for a fully connected network is considered in [23], where a simplified uniform backoff scheme is assumed.

In the paper, we consider the more complex *standard* backoff scheme with the aim of minimizing changes of the existing and widely adopted protocol.

2 IEEE 802.11 DCF: Basic Principles and Enhancements

The basic 802.11 MAC protocol, the Distributed Coordination Function (DCF), works as listen-before-talk scheme based on Carrier Sense Multiple Access (CSMA), with a Collision Avoidance (CA) mechanism to avoid collisions that can be anticipated if terminals are aware of the duration of ongoing transmissions ("virtual carrier sense"). 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 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 and increments the retry counter. The backoff time is randomly and uniformly chosen in the range $(0, W-1)$, W being the contention window. The MAC decrements the backoff value each time the medium are 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 ACK frame from the receiver after SIFS (plus the propagation delay) or it will consider the transmission as failed. If a failure happens, the station reschedules the packet transmission according to the given backoff rules. At the first transmission attempt, W is set equal to a value $CW_{\min}$ called minimum contention window. After each unsuccessful transmission, W is doubled, up to a maximum value $CW_{\max} = 2^m \cdot CW_{\min}$.

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 capable of service differentiation [24]-[25]. In the interest of conciseness, we are interested in gaining insight into one of the building block used to achieve differentiation, i.e. differentiating the minimum contention window sizes according to the priority of each traffic category.

2.1 System Modeling

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

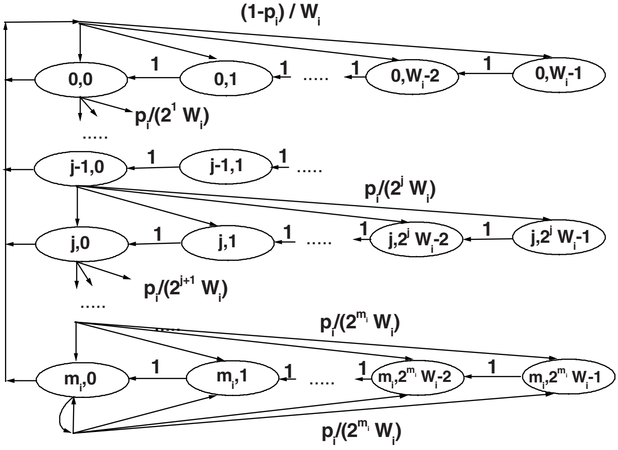

Fig. 1. Markov model of backoff process for type-i traffic

Because our analysis can be easily extended and for the sake of simplicity, only two different types of traffic are considered with n_i traffic flows for traffic of type i $(i=1,2)$. Moreover, it is assumed that each mobile terminal has only one traffic flow. Let $b_i(t)$ be the stochastic process representing the backoff time counter for a given traffic flow with type i . Moreover, let us define for convenience $W_i = CW_{\min,i}$ as the minimum contention window for traffic type i. Let m_i, "maximum backoff stage" be the value such that $CW_{\max,i} = 2^{m_i} \cdot W_i$. Let $s_i(t)$ be the stochastic process representing the backoff stage $(0,1,...,m_i)$ for a given traffic flow with type i.

We use a two-dimensional discrete-time Markov chain to model the behavior of a traffic flow with type i. The states are defined as the combinations of two integers $\{s_i(t), b_i(t)\}$. The Markov chain for type-i traffic flows are shown in Fig.1. All details about the analysis can be found in [10], [18] and [29].

2.2 Throughput Analysis

Let $q_i(j,k)$, $j \in [0, m_i]$ and $k \in [0, 2^j \cdot W_i - 1]$, be the stationary distribution of the chain. It is easy to find that

$$q_i(0,0) = \left(2(1-2p_i)(1-p_i)\right) \Big/ \left((1-2p_i)(W_i+1) + p_i W_i [1-(2p_i)^{m_i}]\right) \tag{1}$$

τ_i is defined as the probability that a station carrying type-i traffic transmits in a randomly chosen slot time. We have

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

With the above probabilities defined, we can express packet collision probabilities p_i as:

$$p_i = 1-(1-\tau_i)^{n_i-1} \prod_{j=1, j\neq i}^{L} (1-\tau_j)^{n_j} \tag{3}$$

After combining equations (2) and (3) and by using Successive Over-Relaxation (SOR) numerical method [26], we can get all the values for p_i and τ_i.

Moreover, we define $Q(i, j)$ as the probability that there are a number i of type-1 stations and a number j of type-2 stations transmitting within a randomly selected slot. Then, we have

$$Q(c_1,c_2) = \binom{n_1}{c_1} \cdot \tau_1^{c_1}(1-\tau_1)^{n_1-c_1} \cdot \binom{n_2}{c_2} \cdot \tau_2^{c_2}(1-\tau_2)^{n_2-c_2} \tag{4}$$

The normalized system throughputs S can be expressed as:

$$\begin{aligned} S &\equiv \frac{\text{Average payload transmitted in a slot time}}{\text{Average length of a slot time}} = S_1 + S_2 \\ &= \frac{Q(1,0)\cdot E[P_{Len,1}] + Q(0,1)\cdot E[P_{Len,2}]}{\left\{ Q(0,0)\cdot\sigma + Q(1,0)\cdot T_{s,1} + Q(0,1)\cdot T_{s,2} + \sum\limits_{0\le c_1 \le n_1, 0\le c_2\le n_2, c_1+c_2\ge 2} Q(c_1,c_2)\cdot T_c(c_1,c_2) \right\}} \\ &\equiv \frac{Q(1,0)\cdot E[P_{Len,1}] + Q(0,1)\cdot E[P_{Len,2}]}{Q(0,0)\cdot\sigma + Q(1,0)\cdot T_{s,1} + Q(0,1)\cdot T_{s,2} + [1-Q(0,0)-Q(1,0)-Q(0,1)]\cdot T_c} \end{aligned} \tag{5}$$

where S_1 and S_2 denote the throughputs contributed by type-1 and type-2 traffic flows, respectively. $E[P_{Len,i}]$ is the average duration to transmit the payload for type-i traffic (the payload size is measured with the time required to transmit it). For simplicity, with the assumption that all packets of type-i traffic have the same fixed size, we have $E[P_{Len,i}] = P_{Len,i}$. σ is the duration of an empty time slot. $T_{s,i}$ is the average time of a slot because of a successful transmission of a packet of a type-i traffic flow. $T_{s,i}$ can be expressed as

$$T_{s,i} = PHY_{header} + MAC_{header} + E[P_{Len,i}] + SIF + \delta + ACK + DIFS + \delta \tag{6}$$

where δ is the propagation delay. $T_c(c_1,c_2)$ is the average time the channel is sensed busy by each station during a collision caused by simultaneous transmissions of c_1 type-1 stations and c_2 type-2 stations. It can be expressed as

$$T_c(c_1,c_2) = PHY_{header} + MAC_{heade} + \max[\theta(c_1)P_{Len,1}, \theta(c_2)P_{Len,2}] + DIFS + \delta \tag{7}$$

where

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

Moreover, from equation (3), we can easily derive

$$(1 - p_1)(1 - \tau_1) = (1 - p_2)(1 - \tau_2) = \prod_{j=1}^{2} (1 - \tau_j)^{n_j} \tag{8}$$

When the minimum contention window size $W_1 >> 1$ and $W_2 >> 1$, the transmission probabilities τ_1 and τ_2 are small, that is, $\tau_1 << 1$ and $\tau_2 << 1$. Therefore, from equation (8), we have the following approximation

$$p_1 \approx p_2 \tag{9}$$

When $W_1 >> 1$, $W_2 >> 1$ and $m_1 \approx m_2$, we have the following approximation based on equation (2)

$$\left(\tau_1 / \tau_2\right) \approx \left(W_2 / W_1\right) \tag{10}$$

From equations (4), (5) and (10), we finally have

$$\frac{s_1}{s_2} \equiv \frac{S_1 / n_1}{S_2 / n_2} = \frac{\frac{\tau_1}{1 - \tau_1} \cdot E[P_{Len,1}]}{\frac{\tau_2}{1 - \tau_2} \cdot E[P_{Len,2}]} \approx \left(\frac{E[P_{Len,1}]}{W_1}\right) \Big/ \left(\frac{E[P_{Len,2}]}{W_2}\right) \tag{11}$$

3 Maximum Throughput Analysis

We are interested in maximizing throughput, while *at the same time* ensuring service differentiation, and the hypothesis in this section is that differentiation is achieved by allocating bandwidth to the individual traffic flow to satisfy a given target ratio $\hat{\alpha} = s_2 / s_1$. For convenience, it is useful to define an additional *differentiation* parameter $\alpha \equiv \left(\frac{\tau_2}{1 - \tau_2}\right) \Big/ \left(\frac{\tau_1}{1 - \tau_1}\right)$. According to equation (11), we have $\alpha = \hat{\alpha} \cdot \frac{E[P_{Len,2}]}{E[P_{Len,1}]}$. In the following we always assume that the probabilities of transmission in a randomly selected slot time satisfy the constraints $0 \le \tau_1 < 1$, $0 \le \tau_2 < 1$.

Theorem 1: Assume that two types of traffic coexist in the system, with n_1 and n_2 numbers of traffic flows, respectively. If one fixes the desired differentiation: $\frac{\tau_2}{1 - \tau_2} = \alpha \cdot \frac{\tau_1}{1 - \tau_1}$ $(\alpha > 0)$, the throughput function $S(\tau_1, \tau_2)$ defined in equation (5) has one and only one optimal operation point $\tau_1^*(\alpha)$ where the maximum throughput is achieved.

Proof:

From equation (5), we have

$$S(\tau_1,\tau_2)=\frac{n_1\frac{\tau_1}{(1-\tau_1)}E[P_{Len,1}]+\alpha n_2\frac{\tau_1}{(1-\tau_1)}E[P_{Len,2}]}{\left\{\begin{array}{l}\sigma+n_1\frac{\tau_1}{1-\tau_1}\cdot T_{s,1}+\alpha n_2\frac{\tau_1}{1-\tau_1}\cdot T_{s,2}\\ +\alpha n_1 n_2\left(\frac{\tau_1}{1-\tau_1}\right)^2\cdot T_c(1,1)+n_1(n_1-1)\left(\frac{\tau_1}{1-\tau_1}\right)^2\cdot T_c(2,0)\\ +\alpha^2 n_2(n_2-1)\left(\frac{\tau_1}{1-\tau_1}\right)^2\cdot T_c(0,2)+\frac{\sum\limits_{c_1+c_2\geq 3,c_1\leq n_1,c_2\leq n_2}^{n_1+n_2}Q(c_1,c_2)T_c(c_1,c_2)}{(1-\tau_1)^{n_1}(1-\tau_2)^{n_2}}\end{array}\right\}} \tag{12}$$

$$=\left(F_1\cdot\chi\right)\Big/\left(\sigma+\sum_{i=1}^{n_1+n_2}G_i\cdot\chi^i\right)\equiv\frac{F(\chi)}{G(\chi)}$$

where $\chi\equiv\frac{\tau_1}{1-\tau_1}$ $(0\leq\chi<+\infty)$, F_1 and G_i $(i=1,2,...,n_1+n_2)$ are constants larger than zero. To determine the optimal operation point, we study the function:

$$\left(\frac{F(\chi)}{G(\chi)}\right)'=\frac{F(\chi)'G(\chi)-F(\chi)G(\chi)'}{G(\chi)^2}=\left(F_1\sigma-F_1\sum_{i=2}^{n_1+n_2}(i-1)G_i\chi^i\right)\Big/G(\chi)^2 \tag{13}$$

The optimal solution χ^* satisfies the following equation:

$$\sum_{i=2}^{n_1+n_2}(i-1)G_i\left(\chi^*\right)^i=\sigma \tag{14}$$

Because $\sigma>0$ and $\sum_{i=2}^{n_1+n_2}(i-1)G_i\chi^i$ is a monotone 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 equation (13), it can be seen that $\left(F(\chi)/G(\chi)\right)'>0$ when $\chi<\chi^*$ and $\left(F(\chi)/G(\chi)\right)'<0$ when $\chi>\chi^*$. Therefore, the throughput function reaches the maximum value when $\frac{\tau_1^*}{1-\tau_1^*}=\chi^*$. Of course the optimal solution varies with the variation of the *differentiation* constant α. Therefore, we denote the optimal solution as $\tau_1^*(\alpha)$. □

By using equation (14), the optimal operation point can be obtained by using a numerical method. However, in order to obtain a much deeper insight into the system performance, it is useful to derive more meaningful and concise approximations of the exact formulas. From equations (12) and (14), we have

$$\frac{\tau_1^*(\alpha)}{1-\tau_1^*(\alpha)} \le \sqrt{\frac{\sigma}{G_2}} = \sqrt{\frac{\sigma}{\alpha n_1 n_2 T_c(1,1) + n_1(n_1-1)T_c(2,0) + \alpha^2 n_2(n_2-1)T_c(0,2)}} \tag{15}$$

It can be seen that, if n_1, n_2, $E[P_{Len,1}]$ and $E[P_{Len,2}]$ are sufficiently large, the optimal operation point $\tau_1^*(\alpha)$ is far less than one (it is also true for $\tau_2^*(\alpha)$). Therefore, it is reasonable to limit the discussions to the case that $\tau_1 << 1$ and $\tau_2 << 1$.

Theorem 2: Assume that two types of traffic coexist in the system with n_1 and n_2 flows, respectively. Moreover, assume that $\frac{\tau_2}{1-\tau_2} = \alpha \cdot \frac{\tau_1}{1-\tau_1}$ $(\alpha > 0)$. If n_1, n_2, $E[P_{Len,1}]$ and $E[P_{Len,2}]$ are sufficiently large so that the optimal operation point $\tau_1^*(\alpha) << 1$, $\tau_2^*(\alpha) << 1$, than the optimal operation point can be approximated as

$$\tau_1^*(\alpha) \approx 1\Big/\left((n_1 + \alpha n_2)\sqrt{T_c^*/2}\right) \equiv \tau_{1\_ap}^*(\alpha) \tag{16}$$

where $T_c^* \equiv T_c/\sigma$. Moreover, if $E[P_{Len,1}] = E[P_{Len,2}] = P_{Len}$, the corresponding achieved maximum throughput can be approximated as

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

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

Proof:

According to Theorem 1, because at the optimal operation point $\tau_1^*(\alpha) << 1$, $\tau_2^*(\alpha) << 1$, we can limit our discussion only to the range of $\tau_1 << 1$, $\tau_2 << 1$. In this case, the relationship $\frac{\tau_2}{1-\tau_2} = \alpha \cdot \frac{\tau_1}{1-\tau_1}$ $(\alpha > 0)$ can be approximated as $\tau_2 \approx \alpha \cdot \tau_1$.

First, if we neglect the case that three or more packets collide with each other at the same time, we have

$$\begin{aligned} T_c &\approx \frac{Q(1,1)\cdot T_c(1,1) + Q(2,0)\cdot T_c(2,0) + Q(0,2)\cdot T_c(0,2)}{Q(1,1)+Q(2,0)+Q(0,2)} \\ &\approx \frac{\alpha n_1 n_2 \cdot T_c(1,1) + n_1(n_1-1)\cdot T_c(2,0) + \alpha^2 n_2(n_2-1)\cdot T_c(0,2)}{\alpha n_1 n_2 + n_1(n_1-1) + \alpha^2 n_2(n_2-1)} \end{aligned} \tag{18}$$

From the above approximation, it can be seen that once $E[P_{Len,1}]$, $E[P_{Len,2}]$, n_1, n_2 and α are given, T_c can be regarded as a constant.

Based on the assumption that $\tau_1 << 1$, $\tau_2 << 1$ and on equation (8), equation (5) can be approximated as follows:

$$S(\tau_1,\tau_2)\approx\frac{n_1\tau_1(1-p_1)E[P_{Len,1}]+n_2\alpha\tau_1(1-p_1)E[P_{Len,2}]}{\left\{\begin{array}{l}(1-\tau_1)(1-p_1)\sigma+n_1\tau_1(1-p_1)T_{s,1}+n_2\alpha\tau_1(1-p_1)T_{s,2}\\+[1-(1-\tau_1)(1-p_1)-n_1\tau_1(1-p_1)-n_2\alpha\tau_1(1-p_1)]T_c\end{array}\right\}}$$
$$\approx\frac{n_1\tau_1E[P_{Len,1}]+n_2\alpha\tau_1E[P_{Len,2}]}{\sigma+n_1\tau_1T_{s,1}+n_2\alpha\tau_1T_{s,2}+[(1-p_1)^{-1}-1-n_1\tau_1-n_2\alpha\tau_1]T_c}\equiv\frac{f(\tau_1)}{g(\tau_1)}\qquad(19)$$

Approximately, the optimal solution must satisfy the following condition

$$f(\tau_1^*)\big/f'(\tau_1^*)=g(\tau_1^*)\big/g'(\tau_1^*)\qquad(20)$$

That is,

$$\tau_1^*=\frac{\sigma+n_1\tau_1^*T_{s,1}+n_2\alpha\tau_1^*T_{s,2}+[\dfrac{1}{1-p_1}-1-n_1\tau_1^*-n_2\alpha\tau_1^*]\cdot T_c}{n_1T_{s,1}+n_2\alpha T_{s,2}+[d(1-p_1)^{-1}\big/d\tau_1\big|_{\tau_1=\tau_1^*}-n_1-n_2\alpha]\cdot T_c}\qquad(21)$$

After some simplifications of the above equation, one obtains

$$(n_1+\alpha n_2)\tau_1^*T_c^*=(1-p_1)\big|_{\tau_1=\tau_1^*}\cdot(1-T_c^*)+T_c^*\approx(1-\tau_1^*)^{n_1}(1-\alpha\tau_1^*)^{n_2}(1-T_c^*)+T_c^*\qquad(22)$$

Because $(1-\alpha\tau_1^*)^{n_2}\approx 1-\alpha n_2\tau_1^*\approx(1-\tau_1^*)^{\alpha n_2}$, the above equation can be further approximated as

$$(n_1+\alpha n_2)\tau_1^*T_c^*=(1-\tau_1^*)^{n_1+\alpha n_2}\cdot(1-T_c^*)+T_c^*\qquad(23)$$

When there is only one type of traffic, equation (23) is actually the same as equation (27) in [10]. By referring to equation (28) in [10], equation (16) can be obtained.

Next, we evaluate the maximum throughput by substituting the approximate optimal solution $\tau_{1\_ap}^*(\alpha)$ into equation (5).

$$S_{\max}\approx\frac{n_1\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}E[P_{Len,1}]+\alpha n_2\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}E[P_{Len,2}]}{\left\{\begin{array}{l}(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}\sigma+n_1\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}T_{s,1}\\+\alpha n_2\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}T_{s,2}+[1-(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}\\-n_1\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}-\alpha n_2\tau_{1\_ap}^*(1-\tau_{1\_ap}^*)^{n_1+\alpha n_2}]\cdot T_c\end{array}\right\}}\qquad(24)$$

Because n_1 and n_2 are assumed sufficiently large, we have the following approximation:

$$\left(1-((n_1+\alpha n_2)K)^{-1}\right)^{n_1+\alpha n_2}\approx e^{-1/K}\qquad(25)$$

Moreover, we assume $E[P_{Len,1}]=E[P_{Len,2}]=P_{Len}$ and therefore $T_{s,1}=T_{s,2}=T_s$, then equation (24) can be further approximated as equation (17). □

Deduction 1: Assume that $L\ge 1$ types of traffic coexist in the system, with numbers of type-i traffic flows n_i $(i=1,2,...,L)$. Moreover, assume that $\dfrac{\tau_i}{1-\tau_i}=\alpha_i\cdot\dfrac{\tau_1}{1-\tau_1}$ $(\alpha_i>0,i=1,2,...,L,\alpha_1\equiv 1)$. If n_i $(i=1,2,...,L)$ and $E[P_{Len,i}]$ $(i=1,2,...,L)$ are sufficiently large so that the optimal operation point $\tau_i^*(\alpha_1,...,\alpha_L)<<1$ $(i=1,2,...,L)$, then the optimal operation point can be approximated as

$$\tau_1^*(\alpha_2,...,\alpha_L) \approx 1 \Bigg/ \left(\sum_{j=1}^{L} \alpha_j n_j \sqrt{\frac{T_c^*}{2}} \right) \equiv \tau_{1\_ap}^*(\alpha_2,...,\alpha_L) \tag{26}$$

where $T_c^* \equiv T_c / \sigma$. Moreover, if $E[P_{Len,1}] = ... = E[P_{Len,L}] = P_{Len}$, the corresponding achieved maximum throughput can be approximated as

$$S_{\max} \approx P_{Len} \Big/ \left(T_s + \sigma K + T_c[K(e^{1/K} - 1) - 1] \right) \tag{27}$$

where $K \equiv \sqrt{T_c^* / 2}$. □

Compared with the equation (31) in [10], we find that *the maximum throughput achieved is exactly the same no matter how many different types of traffic flows coexisting in the system.*

Deduction 2: Assume that there are $L \geq 1$ types of traffic coexisting in the system with n_i $(i = 1,2,...,L)$ traffic flows. Moreover, assume that $\frac{\tau_i}{1-\tau_i} = \alpha_i \cdot \frac{\tau_1}{1-\tau_1}$ $(\alpha_i > 0, i = 1,2,...,L, \alpha_1 \equiv 1)$. If n_i $(i = 1,2,...,L)$ and $E[P_{Len,i}]$ $(i = 1,2,...,L)$ are sufficiently large so that the optimal operation point $\tau_i^*(\alpha_1,...,\alpha_L) << 1$ $(i = 1,2,...,L)$, then the system operates close to the optimal operation point if and only if the packet collision rate is approximately equal to $1 - e^{-1/K}$ ($K \equiv \sqrt{T_c^* / 2}$). □

The above equation can be used to check if the system works close to the optimal operation point.

4 Validation of Approximations

In this section, we validate the approximated results obtained in the former section by using a numerical method. The parameters for the system are summarized in Table 1, based on IEEE 802.11b.

In the first example, we compare the exact optimal operation points τ_1^* numerically obtained from equation (5) with the approximated optimal operation

Table 1. System Parameters

MAC Header	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

points $\tau_{1\_ap}^{*}$ obtained from equation (16). In the example, we set other parameters as: $\frac{\tau_2}{1-\tau_2}=\alpha\frac{\tau_1}{1-\tau_1}$, $\frac{n_2}{n_1}=2$, $P_{Len,1}=P_{Len,2}=2000$ *bytes*, and $m_1=m_2=8$. In Fig. 2, the comparison results of optimal operation points are shown versus the number of type-1 traffic flows n_1. Two cases are shown in the figure: one is for the case that $\alpha=0.1$ and the other is $\alpha=10$. 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 case of $\alpha=0.1$ and that of $\alpha=10$ show that good estimation accuracy can be obtained as long as the estimated optimal operation point are far less than one.

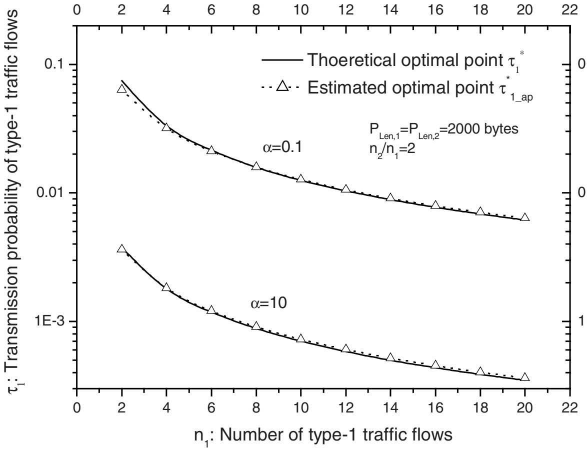

Fig. 2. Comparisons between theoretical optimal operation points and estimated ones

After verifying the accuracy of the estimation for 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 exact optimal operational point and its corresponding approximated one into equation (5) respectively. The comparison results are given in Table 2. From the Table, it can be seen that the estimated maximum throughput $S_{\max\_ap}$ accord with the corresponding theoretical value $S_{\max}$ very well. Moreover, in the Table, we show the evaluated maximum throughput obtained from equation (17). It can be regarded as the limiting value for the maximum throughput when $n_1 \to \infty$.

5 An Adaptive Scheme to Achieve Maximum Throughput and Service Differentiation

For the implementation of real-world systems, in addition to the existence of an optimal operation point, one is interested in methods to reach the point and to maintain a dynamic system close to the optimal point. In the following part, we present two schemes for this purpose.

Table 2. Comparisons between theoretical maximum throughput and estimated ones

	α=0.1		α=10	
n_1	S_{max}	S_{max_ap}	S_{max}	S_{max_ap}
6	0.66521	0.66518	0.66323	0.66322
8	0.66383	0.66381	0.66237	0.66235
10	0.66301	0.66299	0.66187	0.66183
12	0.66248	0.66245	0.66153	0.66148
14	0.66210	0.66206	0.66129	0.66123
16	0.66181	0.66177	0.66111	0.66105
18	0.66159	0.66155	0.66097	0.66091
20	0.66142	0.66137	0.66086	0.66079
∞	0.65976			

System parameters: $P_{Len,1} = P_{Len,2} = 2000$ bytes, $n_2 = 2n_1$, $m_1=m_2=8$

5.1 Basic Adaptive Scheme

Based on equation (11), to achieve a certain target service differentiation $\hat{\alpha} = s_2/s_1$, we can adjust the ratio $\left(\frac{\tau_2}{1-\tau_2}\right)\Big/\left(\frac{\tau_1}{1-\tau_1}\right)$ to be $\alpha = \hat{\alpha} \cdot \frac{E[P_{Len,1}]}{E[P_{Len,2}]}$. In this case, if the optimal operation point $\tau_1^* << 1$ and $\tau_2^* << 1$, from Theorem 2, they can be approximated as $\tau_1^* \approx 1\Big/\left((n_1 + \alpha n_2)\cdot\sqrt{T_c^*/2}\right)$ and $\tau_2^* \approx 1\Big/\left((\frac{n_1}{\alpha} + n_2)\cdot\sqrt{T_c^*/2}\right)$, respectively. Therefore, if $E_1 \equiv n_1 + \alpha n_2$ and $E_2 \equiv \frac{n_1}{\alpha} + n_2 = \frac{E_1}{\alpha}$ are known and the packet transmission probabilities of traffic flows are equal to $\tau_1^* \approx 1\Big/\left(E_1 \cdot\sqrt{T_c^*/2}\right)$ and $\tau_2^* \approx 1\Big/\left(E_2 \cdot\sqrt{T_c^*/2}\right)$ respectively, the system operates *almost* at the optimal point and the service differentiation achieved can be approximated as $\frac{s_2}{s_1} \approx \hat{\alpha} = \alpha \cdot \frac{E[P_{Len,2}]}{E[P_{Len,1}]}$.

Assuming that n_1, n_2 and α are known, the problem is how to make packet transmission probabilities τ_1 and τ_2 reach their corresponding approximate optimal values $\tau^*_{1\_ap}$ and $\tau^*_{2\_ap}$. First, each station can evaluate the average frame collision length T_c^* at run-time. Next, it calculates the target optimal packet transmission probabilities ${\tau_{1\_ap}}^*$ or ${\tau_{2\_ap}}^*$ based on Theorem 2, and the approximate packet collision rate $p^*_{\_ap}$ corresponding to the optimal operation point by using Deduction 2. Then, by substituting ${\tau_{1\_ap}}^*$, ${\tau_{2\_ap}}^*$ and $p^*_{\_ap}$ into equation (2), one can obtain the approximate optimal minimum contention window size $W^*_{1\_ap}$ and $W^*_{2\_ap}$. Finally, $W^*_{1\_ap}$ and $W^*_{2\_ap}$ are used to adjust the current minimum contention window size $Current\_W_1$ and $Current\_W_2$ as follows:

$$Current\_W_i = \beta \cdot Current\_W_i + (1-\beta) \cdot W^*_{i\_ap} \tag{28}$$

where $i = 1,2$, and $\beta \in [0,1]$ is a smoothing factor, which determines the convergence speed of the scheme.

We simulated the above scheme to verify its performance. In the simulation, it is assumed that $n_1 = 10$, $n_2 = 20$ are known. In this case, no central controller is needed. Parameter α is set as 0.2. The frame lengths of both traffic types are equal. Both traffic flows begin their minimum contention window size from 512.

Table 3 shows the comparison between the theoretical maximum throughput $S_{\max}$ and the actual throughput S and the service differentiation s_1/s_2 achieved by using the basic adaptive scheme. It can be seen that the proposed adaptive scheme can achieve the maximum throughput and at the same time the target service differentiation performance.

Table 3. Comparisons between theoretical maximum throughput and simulated ones

P_{Len} (bytes)	S_{max}	S	s_1/s_2
500	0.36199	0.36235	5.01728
700	0.43628	0.43588	5.02386
900	0.49298	0.49316	5.07283
1100	0.53786	0.53677	4.98651
1300	0.57437	0.57460	5.05422
1500	0.60471	0.60646	5.02912
1700	0.63038	0.63139	4.97099
1900	0.65241	0.65302	5.10741
2100	0.67155	0.67105	5.11010

n_1=10, n_2 = 20, 1/ = 5.0,   = 0.8, m_1=m_2=8

In the basic adaptive scheme, it is assumed that n_1, n_2 and α are known (hence E_1 and E_2 are known). The optimal operation point ${\tau_{1\_ap}}^*$, ${\tau_{2\_ap}}^*$ are mainly determined by the value of E_1 and E_2. However, extensive simulations show that the sensitivity of the achieved throughput to changes of E_1 is small, when the differentiation parameter α is fixed. To some extent, the system can achieve optimal performance by using the basic adaptive scheme even the actual number of traffic flows are different from the assumed ones. This is because the throughput function in equation (5) is very smooth with the variation of τ_1. However, for large deviations of E_1 from the assumed value, the achieved throughput deteriorates.

5.2 A Centralized Adaptive Scheme

A centralized version of the adaptive scheme uses a central controller (CC) is proposed in this section. Let us note that a centralized network control can be assumed in a *hot spot* scenario, with the need of identifying users, accounting and billing, managing and supporting QoS, possibly also through pricing and call admission control (CAC) [27]. In our scheme, the CC itself carries traffic flows for transmission (we assume 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 and E_2 at run time. If the detected value of E_1 and E_2 are sufficiently far from the current estimates, the CC broadcasts the new estimates. In order to maintain the target service differentiation between different traffic flows, the CC also broadcasts the target differentiation ratio. After receiving the new values, all mobile terminals in the current basic service set (BSS) modify their memorized values of E_1 and E_2 and use the adaptive scheme described previously.

To keep track of the number of active mobile stations, the CC monitors the traffic and evaluates the real-time values of E_1 and E_2 as follows. In the case that $\tau_1 << 1$, $\tau_2 << 1$, $\tau_2 = \alpha\tau_1$ and by using equation (8), one has

$$(1-p_1) \approx (1-\tau_1)^{n_1+\alpha n_2} = (1-\tau_1)^{E_1} \tag{29}$$

From above equation, one estimates E_1 as

$$\hat{E}_1 = \log(1-p_1)/\log(1-\tau_1) \tag{30}$$

where the packet collision rate p_1 can be easily evaluated at run-time. An efficient way to evaluate the run-time packet collision rate is proposed in [28]. τ_1 is obtained by substituting the estimated p_1 and the current minimum contention window size $Current\_W_1$ into equation (2). After obtaining $\hat{E}_1$, it is averaged as $\overline{E_1}$ and compared with the $Current\_E_1$, which is the current memorized value for E_1. $\overline{E_1}$ can be expressed as

$$\overline{E_1} = \beta \cdot \overline{E_1} + (1-\beta) \cdot \hat{E}_1 \tag{31}$$

Table 4. Performance of the modified adaptive scheme

n_1,n_2	S_{max}	S	s_1/s_2
2,4	0.67338	0.66721	5.67252
5,10	0.66486	0.66508	5.35558
10,20	0.66230	0.66184	4.96943
20,40	0.66107	0.66238	5.00814
30,60	0.66066	0.65910	4.93129
50,100	0.66035	0.65292	4.83726

P_{Len}=2000 bytes, 1/ = 5.0, = 0.8, m_1=m_2=8

If $\overline{E_1}$ is less than $Current\_E_1 \cdot \gamma$ ($0 < \gamma < 1$) during the past $k_t \geq 1$ comparisons, the $Current\_E_1$ will be set as $\overline{E_1}$. If $\overline{E_1}$ is larger than $Current\_E_1 / \gamma$ ($0 < \gamma < 1$) during the past $k_t \geq 1$ comparisons, the $Current\_E_1$ will be set as $\overline{E_1}$. $Current\_E_2$ is simply obtained as $Current\_E_1 / \alpha$. In the scheme, if γ is set to be 0, the improved scheme is actually the same as the basic scheme. On the other hand, if γ is very close to 1, the CC will modify E_1 and E_2 too often, which proves to be unnecessary according to the former discussions about the sensitivities of achieved throughput to the number of traffic flows. Therefore, parameters γ and k_t should be carefully chosen to improve the performance of the system and to minimize the control overhead.

The performance of the improved scheme is verified by simulation. In the simulation, a station carrying type-1 traffic flow serves as the CC. γ and k_t are set to be 0.5 and 10, respectively. If the CC decides to broadcast new values for E_1 and E_2, it generates a special management frame and gains access to the channel by using the highest medium access priority (PIFS) to ensure the new values can be received as soon as possible. Table 4 shows the performance of the centralized adaptive scheme. We can see that the achieved throughput S is now close to the corresponding maximum throughput $S_{\max}$ for all the cases, which is caused by the ability to adapt to dynamically changing values of E_1 and E_2. Moreover, that service differentiation ratio s_1 / s_2 is kept approximately constant.

6 Conclusions

In this paper, we use a model of a wireless LAN based on the *standard* IEEE 802.11 MAC with a simple extension for service differentiation and derive approximations to get simpler but more meaningful relationships among the different parameters. We successfully derive the best operation point where the maximum throughput can be achieved and demonstrate its uniqueness. In addition we propose simple rules to decide if the system works under the optimal state. The other contribution of the paper is the proposal of two adaptive schemes (one distributed and the other one centralized) to lead and maintain the system close to the optimal operation point while at the same time guaranteeing target service differentiation between different traffic types.

References

1. Y. Cheng and W. H. Zhuang, “DiffServ resource allocation for fast handoff in wireless mobile Internet,” IEEE Communications Magazine, vol. 40, no. 5, 2002, pp. 130–136.
2. R. Braden, D. Clark and S. Shenker, “Integrated services in the Internet architecture: an overview,” IETF RFC 1633, Jun. 1994.
3. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, “An architecture for differential services,” IETF RFC 2475, Dec. 1998.
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 Networks and Applications, vol. 2, pp. 55–67, 1997.
6. H. S. Chhaya and S. Gupta, “Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol,” Wireless Networks, vol. 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,” Proceedings of IEEE PIMRC, Taipei, Taiwan, Oct. 1996, pp.392–396.
8. F. Cali, M. Conti, and E. Gregori, “IEEE 802.11 wireless LAN: Capacity analysis and protocol enhancement,” Proceedings of INFOCOM’98, San Francisco, CA, March 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,” Proceedings of IEEE PIMRC, Taipei, Taiwan, Oct. 1996, pp. 407–411.
10. G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” IEEE Journal on Selected Areas In Communications, vol. 18, no. 3, March 2000.
11. Y. C. Tay and K. C. Chua, “A Capacity Analysis for the IEEE 802.11 MAC Protocol,” Wireless Networks, 7, 2001, pp. 159–171.
12. J. L. Sobrinho and A. S. Krishnakumar, “Distributed multiple access procedures to provide voice communications over IEEE 802.11 wireless networks,” Proceedings GLOBECOM 1996, pp. 1689–1694.
13. J. Deng and R.S. Chang, “A priority scheme for IEEE 802.11 DCF access method,” IEICE Transactions in Communications, vol. 82-B, no. 1, Jan 1999, pp. 96–102.
14. A. Veres, A. T. Campbell, M. Barry and L. H. Sun, “Supporting Service Differentiation in Wirelss Packet Networks Using Distributed Control,” IEEE Journal on Selected Areas In Communications, vol. 19, no. 10, Oct 2001, pp. 2081–2093.
15. I. Aad and C. Castelluccia, “Differentiation Mechanisms for IEEE 802.11,” Proceedings of IEEE Inforcom 2001, pp. 209–218.
16. S. Mangold, S. Choi, P. May, O. Klein, G. Hietz and L. Stibor, “IEEE 802.11e wireless lan for quality of service,” Proceedings of the European Wireless, Feb 2002.
17. Z. Jun, G. Zihua, Z. Qian and Z. Wenwu, “Performance Study of MAC for Service Differentiation in IEEE 802.11,” Proceedings of the GLOBECOM '02, IEEE , Volume: 1 , Nov 17-21, 2002 pp. 778–782
18. Bo LI, Roberto Battiti, “Supporting Service Differentiation with Enhancements of the IEEE 802.11 MAC Protocol: Models and Analysis” Technical Report of Department of Computer Science and Telecommunications of University of Trento, no. DIT-03-024, available at http://dit.unitn.it/research/publications/techRep?id=418

19. F. Cali, M. Conti and E. Gregori "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," IEEE/ACM Transactions on Networking, vol. 8, No. 6, Dec 2000, pp. 785–799.
20. L. Romdhani, Q. Ni, and T. Turletti, "AEDCF: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad-Hoc Networks," INRIA Technical Report. http://www.inria.fr/rrrt/rr-4544.html
21. T. Ozugur, M. Naghshineh, P. Kermani, C. Michael, B. Rezvani and J. A. Copeland, "Balanced media access methods for wireless networks," in Proc. ACM MobiCom'98, Dallas, TX, Oct. 1998, pp.21–32.
22. N. H. Vaidya, P. Bahl and S. Gupta, "Distributed fair scheduling in a wireless Lan," ACM Mobicom'2000. http://research.microsoft.com/users/bahl/papers/pdf/mobiCom2000.pdf
23. D. Qiao and K. G. Shin, "Achieving Efficient Channel Utilization and Weighted Fairness for Data Communications in IEEE 802.11 WLAN under the DCF," Quality of Service, 2002. Tenth IEEE International Workshop, 2002, Page(s): 227–236
24. M. Benveniste, G. Chesson, M. Hoehen, A. Singla, H. Teunissen, and M.Wentink, "EDCF proposed draft text," IEEE working document 802.11-01/131r1, March 2001.
25. A. Lindgren, A. Almquist, and O. Schelen, "Evaluation of quality of service schemes for IEEE 802.11 wireless LANs," Proceedings of IEEE Conference on Local Computer Networks (LCN 2001), November 15-16, 2001, pp. 348–351.
26. G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, Wiley-Interscience, 1998, pp. 140–144.
27. Roberto Battiti, Marco Conti, Enrico Gregori, Mikalai Sabel, "Price-based Congestion-Control in Wi-Fi Hot Spots," Proceedings of WiOpt'03 March 3-5, 2003, INRIA Sophia-Antipolis, France, pp. 91–100.
28. G. Bianchi and I. Tinnirello, "Kalman Filter Estimation of the number of Competing Terminals in an IEEE 802.11 Network," IEEE INFOCOM 2003.
29. Bo Li, Roberto Battiti, "Performance Analysis of An Enhanced IEEE 802.11 Distributed Coordination Function Supporting Service Differentiation," QoFIS (International Workshop on Quality of Future Internet Services) 2003, Sweden, Springer Lecture Notes on Computer Science LNCS volume 2811, pp. 152–161.