Research Article

On the Optimization of the IEEE 802.11 DCF: A Cross-Layer Perspective

Massimiliano Laddomada¹ and Fabio Mesiti²

¹ Electrical Engineering Department, Texas A&M University-Texarkana, Texarkana, TX 75505, USA
² DELEN, Politecnico di Torino, 10129 Torino, Italy

Correspondence should be addressed to Massimiliano Laddomada, mladdomada@tamut.edu

Received 5 May 2010; Accepted 20 October 2010

1. Introduction

DCF represents the main access mechanism at the medium access control (MAC) layer [1] of the IEEE 802.11 series of standards, and it is based on carrier sense multiple access with collision avoidance (CSMA/CA).

Many papers, following the seminal work by Bianchi [2], have addressed the problem of modeling, as well as optimizing, the DCF in a variety of traffic load models and transmission channel conditions.

Let us provide a survey of the recent literature related to the problem addressed in this paper. Papers [3–7] model the influence of real channel conditions on the throughput of the DCF operating in saturated traffic conditions. Paper [3] investigates the saturation throughput of IEEE 802.11 in presence of nonideal transmission channel and capture effects.

The behavior of the DCF of IEEE 802.11 WLANs in unsaturated traffic conditions has been analyzed in [8–13], whereby the authors proposed various bidimensional Markov models for unsaturated traffic conditions, extending the basic bidimensional model proposed by Bianchi [2]. In [14], the authors look at the impact of channel-induced errors and of the received signal-to-noise ratio (SNR) on the achievable throughput in a system with rate adaptation, whereby the transmission rate of the terminal is modified depending on either direct, or indirect measurements of the link quality.

The effect of the contention window size on the performance of the DCF has been investigated in [15–22] upon assuming a variety of transmission scenarios. We invite the interested reader to refer to the works [15–22] and references therein.

In [15], the authors proposed a framework to derive the average size of the contention window that maximizes the network throughput in saturated traffic conditions. Moreover, a distributed algorithm aimed at tuning the backoff algorithm of each contending station was proposed. In [16] (see also [17]), the authors investigated by simulation four schemes to reduce the contention window size of...
the contending stations in a saturated network. Moreover, the service differentiation mechanism proposed in the IEEE802.11e standard was used to enhance the offered QoS. Paper [18] provided a theoretical framework to optimize the single-user DCF saturated throughput by selecting the transmitted bit rate and payload size as a function of some specific fading channels. However, no packet error constraint was imposed on the transmitted packets. In [19] the authors focused on the optimization of the single-user DCF saturated throughput by varying the PHY layer data rate and the payload size given a packet error rate constraint.

Paper [20] proposed a constant-window backoff scheme in order to guarantee fairness among the contending stations in the investigated network. Comparisons with the binary exponential backoff used in IEEE 802.11 DCF were also given. In [22] the authors proposed a control theoretic approach to adjust the CW of the contending stations based on the average number of consecutive idle slots between two transmissions. Finally, in [21] the authors presented some European projects where cross-layer optimization techniques were under investigation.

As a starting point for the derivations that follow, we adopt the bidimensional model proposed in a companion paper [12] (see also [13]). Briefly, in [12] the authors extended the saturated Bianchi’s model by introducing a new idle state, not present in the original Bianchi’s model, accounting for the case in which the station queue is empty. The Markov chain of the proposed contention model was solved for obtaining the stationary probabilities, along with the probability τ that a station starts transmitting in a randomly chosen time slot.

Compared to the works proposed in the literature, the novel contributions of this paper can be summarized as follow. Firstly, we investigate the behavior of the aggregate throughput as a function of the traffic load λ of the contending stations, and note that the aggregate throughput S(λ) presents two distinct operating regions identified, respectively, by below link capacity (BLC) region and link capacity (LC) region. Derived in closed-form in this work, and identified by S_m throughout the paper, the link capacity of the considered scenario corresponds to the maximum throughput that the network can achieve when the contending stations transmit with a proper set of network parameters. We show that the network operates in the BLC region when the actual value of the packet rate λ is less than a critical value λ_c, theoretically derived in this work.

The second part of this paper is focused on the optimization of the DCF throughput under variable loading conditions. We propose a cross-layer algorithm whose main aim is to allow the network to operate as close as possible to the link capacity S_m. The proposed optimization algorithm relies on a number of insights derived by the behavior of the aggregate throughput S(λ) as a function of the traffic load λ, and aims at choosing either an appropriate value of the minimum contention window W_0, or a proper size of the transmitted packets depending on the network operating region. The optimization algorithm is dynamic in that it has to be reiterated in order to follow the variations of the network parameters. Finally, we derive a simple model of the optimized throughput, which is useful for predicting the aggregate throughput without resorting to simulation.

The rest of the paper is organized as follows. Section 2 briefly presents the Markov model at the very basis of the proposed optimization framework, whereas Section 3 investigates the behavior of the aggregate throughput as a function of the traffic load λ. The proposed optimization algorithm is discussed in Section 4. Section 5 presents simulation results of the proposed technique applied to a sample network scenario, while Section 6 draws the conclusions.

### 2. Markov Modeling

The bidimensional Markov Process of the contention model proposed in [12], and shown in Figure 1 for completeness, governs the behavior of each contending station through a series of states indexed by the pair (i,k), for all i ∈ [0,m], k ∈ [0,W_i−1], whereby i identifies the backoff stage. On the other hand, the index k, which belongs to the set [0,W_i−1], identifies the backoff counter. By this setup, the size of the ith contention window is W_i = 2^iW_0, for all i ∈ [1,m], while W_0 is the minimum size of the contention window.

An idle state, identified by I in Figure 1, is introduced in order to account for the scenario in which after a successful transmission there are no packets to be transmitted, as well as for the situation in which the packet queue is empty and the station is waiting for a new packet arrival.

The proposed Markov model accounts for packet errors due to imperfect channel conditions, by defining an equivalent probability of failed transmission, identified by P_{eq}, which considers the need for a new contention due either to packet collisions (P_{col}) or to channel errors (P_e) on the transmitted packets, that is,

\[ P_{eq} = P_{col} + P_e - P_e \cdot P_{col}. \]  

It is assumed that at each transmission attempt each station encounters a constant and independent probability of failed transmission, independently from the number of retransmissions already suffered.

The Markov Process depicted in Figure 1 is governed by the following transition probabilities (P_{i,k,j,n} is short for P{s(t+1) = i, b(t+1) = k | s(t) = j, b(t) = n,}):

\[
\begin{align*}
P_{i,k|i,k+1} & = 1, \quad k \in [0,W_i-2], \quad i \in [0,m], \\
P_{0,k|i,0} & = \frac{q(1-P_{eq})}{W_0}, \quad k \in [0,W_0-1], \quad i \in [0,m], \\
P_{i,k|i-1,0} & = \frac{P_{eq}}{W_i}, \quad k \in [0,W_i-1], \quad i \in [1,m], \\
P_{m,k|m,0} & = \frac{P_{eq}}{W_m}, \quad k \in [0,W_m-1], \\
P_{i|i,0} & = (1-q)(1-P_{eq}), \quad i \in [0,m], \\
P_{0,k|i} & = \frac{q}{W_0}, \quad k \in [0,W_0-1], \\
P_{i|i} & = 1-q.
\end{align*}
\]
The first equation in (2) states that, at the beginning of each slot time, the backoff time is decremented. The second equation accounts for the fact that after a successful transmission, a new packet transmission starts with backoff stage 0 with probability $q$, in case there is a new packet in the buffer to be transmitted. Third and fourth equations deal with unsuccessful transmissions and the need to reschedule a new contention stage. The fifth equation deals with the practical situation in which after a successful transmission, the buffer of the station is empty, and as a consequence, the station transits in the idle state $I$ waiting for a new packet arrival. The sixth equation models the situation in which a new packet arrives in the station buffer, and a new backoff procedure is scheduled. Finally, the seventh equation models the situation in which there are no packets to be transmitted and the station is in the idle state.

The stationary distribution of the Markov Model in Figure 1 is employed to compute $\tau$, the probability that a station starts a transmission in a randomly chosen time slot. First of all, observe that a packet transmission takes place when a station goes in one of the states $b_{i,0}$, for all $i \in \{0, m\}$. Therefore, the probability $\tau$ can be evaluated by adding up the probabilities $b_{i,0}$ over the set $i \in \{0, \ldots, m\}$. Upon imposing the normalization condition (The probabilities $b_{j,k}$ must add up to 1 for all $k \in \{1, \ldots, W_i - 1\}$, and for all $i \in \{0, m\}$) on the Markov model in Figure 1, and expressing the probabilities $b_{j,k}$, for all $k \in \{1, \ldots, W_i - 1\}$, as a function of the probabilities $b_{i,0}$, the probability $\tau$ can be rewritten as follows:

$$2\left(1 - 2P_{eq}\right)q \left[q \left[W_0 + 1 \left(1 - 2P_{eq}\right)\right] + W_0 P_{eq} \left(1 - \left(2P_{eq}\right)^m\right)\right] + 2(1 - q) \left(1 - P_{eq}\right) \left(1 - 2P_{eq}\right)^{-1}.$$  

As in Bianchi’s work [2], we assume that: (1) the probability $\tau$ is constant across all time slots; (2) the probability $P_{col}$ is constant and independent of the number of collisions already suffered.

Given $\tau$, the probability of collision $P_{col}$, can be defined as follows:

$$P_{col} = 1 - (1 - \tau)^{N-1}. \quad (4)$$

Finally, the term $q$ in (3) is the probability of having at least one packet waiting for transmission in the station queue after an average slot duration.

Let us spend few words on the evaluation of $q$. Upon assuming that the packet interarrival times are exponentially distributed with mean ($\lambda$, which is measured in pkt/s, represents the rate at which the packets arrive into the station queue from the upper layers.) $1/\lambda$, the probability $q$ can be well approximated by the following relation in a scenario

---

**Figure 1:** Markov chain for the contention model accounting for unsaturated traffic conditions. The model considers $m$ backoff stages. $P_{eq}$ is the probability of failed transmission, accounting for both collisions and channel errors affecting the transmitted packets.
where the contending stations employ queues of small sizes [9]:

\[
q = 1 - e^{-\lambda E[S_t]} \approx \lambda \cdot E[S_t] \bigg|_{\lambda \to 0^+},
\]

\[
E[S_t] = (1 - P_t)\sigma + P_t(1 - P_e)T_c + P_tP_e(1 - \tau)T_c,
\]

where \( E[S_t] \), the expected time per slot, is useful to relate the states of the Markov chain to the actual time spent in each state. As a note aside, notice that \( e^{-\lambda E[S_t]} \) in (5) corresponds to the probability that zero packets are received from the upper layers when the packet interarrival times are exponentially distributed.

The other terms involved in \( E[S_t] \) are defined as follows [2, 12]: \( \sigma \) is the probability that there is at least one transmission in the considered time slot; \( P_t \) is the probability that the considered station transmits with probability \( \tau \); \( P_e \) is the conditional probability that a packet transmission occurring on the channel is successful; \( T_c, T_e \), and \( T_e \) are, respectively, the average times a channel is sensed busy due to a collision, the transmission time during which the data frame is affected by channel errors, and the average time of successful data frame transmission. These times are defined as follows:

\[
T_c = H + E[\text{PL}] + \text{ACKtimeout},
\]

\[
T_e = H + E[\text{PL}] + \text{ACKtimeout},
\]

\[
T_e = H + E[\text{PL}] + SIFS + \tau_p + ACK + DIFS + \tau_p,
\]

where \( \tau_p \) is the propagation delay, and \( H \) accounts for the PHY and MAC header durations. The other times are noticed in Table 1.

### 3. Throughput Analysis

The computation of the normalized system throughput relies on the numerical solution of the nonlinear system obtained by jointly solving (1) and (3). The solution of the system, which corresponds to the values of \( \tau \) and \( P_{eq} \), is used for the computation of the normalized system throughput, that is, the fraction of the time during which the channel is used to successfully transmit payload bits:

\[
S = P_t \cdot P_s \cdot \frac{(1 - P_e)E[\text{PL}]}{E[S_t]},
\]

whereby \( E[\text{PL}] \) is the average packet payload length, and \( P_t \cdot P_s \) can be rewritten as follows [2]:

\[
P_t \cdot P_s = N\tau(1 - \tau)^{N-1}.
\]

In order to gain insights on the behavior of the aggregate throughput \( S \), let us investigate the theoretical behavior of (7) as a function of the packet rate \( \lambda \) for two different values of the packet error probability \( P_e \) and minimum contention window \( W_0 = 32 \), in a scenario with \( N = 10 \) contending stations transmitting at the bit rate 1 Mbps.

<table>
<thead>
<tr>
<th>Table 1: Typical network parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAC header</strong></td>
</tr>
<tr>
<td><strong>PHY header</strong></td>
</tr>
<tr>
<td><strong>ACK</strong></td>
</tr>
<tr>
<td><strong>τ_p</strong></td>
</tr>
<tr>
<td><strong>W_0</strong></td>
</tr>
<tr>
<td><strong>m</strong></td>
</tr>
</tbody>
</table>

The two rightmost subplots of Figure 2 show the theoretical behavior of the throughput in (7), as well as simulation results obtained with NS2 by employing the typical MAC layer parameters for IEEE802.11b given in Table 1 [1]. Other parameters are noticed in the label of the figure. Let us spend a few words about the simulation setup in ns-2. In order to account for imperfect channel transmissions, the channel model is implemented using the suggestions proposed in [23], where the outcomes of a binary, equiprobable random variable are used to establish whether each packet is received erroneously. In other words, the random variable is equal to 1 with probability \( P_e \) (erroneous transmission), and 0 with probability \( 1 - P_e \). We notice in passing that the theoretical model developed is independent on the specific propagation channel. The only parameter needed is \( P_e \), which can be appropriately linked to the specific wireless propagation channel upon specifying a threshold of the signal-to-noise ratio allowing perfect reception at the receiver [24].

We adopted the patch NOAH (NO Ad-Hoc), available on the authors’ website http://icapeople.epfl.ch/ widmer/uwb/ns-2/noah/, for emulating a wireless network in infrastructure mode. The employed traffic model is implemented by generating an exponentially distributed random variable with expected value \( 1/\lambda \), in accordance to the theoretical model developed in Section 2.

Let us focus on the curves noticed in the two rightmost subplots in Figure 2. Basically, there are two different operating regions of the throughput in (7). As \( \lambda \to 0^+ \), that is, all the contending stations approach unloaded traffic conditions, the throughput can be approximated as a straight line passing through the point \( (S, \lambda) = (0, 0) \):

\[
S(\lambda) = N \cdot E[\text{PL}] \cdot \lambda.
\]

This relation follows from the theoretical throughput noticed in (7) upon approximating the probabilities \( P_t \) and \( P_e \) in the limit \( \lambda \to 0^+ \).

Indeed, as \( \lambda \to 0^+ \), the first relation in (5) yields \( q \approx \lambda \cdot E[S_t] \), whereas the probability \( \tau \) in (3) can be well approximated by \( \tau \approx q/(1 - P_{eq}) \). Since \( P_{col} \to 0 \) as \( \lambda \to 0^+ \) (because \( \tau \to 0 \)), it is \( P_{eq} \to P_e \). Furthermore, as \( \lambda \to 0^+ \), (8) can be approximated by the following relation:

\[
P_t \cdot P_s \approx N\tau \frac{q}{1 - P_e}.
\]
Upon substituting (10) in (7), the theoretical throughput can be well approximated as follows:

\[
S(\lambda)_{\lambda=0^+} = P_t \cdot P_s \cdot (1 - P_t) E[PL] \cdot \frac{q}{E[S_{\text{st}}]} \approx NE[PL] \cdot q \approx N \cdot E[PL] \cdot \lambda. \tag{11}
\]

This linear model is depicted in the subplots of Figure 2 overimposed to both theoretical and simulated results. The key observation from this result is that the aggregate throughput produced by \( N \) stations approaching unloaded traffic conditions, is only dependent on the number of stations, as well as on the packet size. No other network parameters affect the aggregate throughput in this operating region. Moreover, the results depicted in Figure 2 denote that the derived linear model is valid up to a critical value of...
The critical value of $\lambda$ acts as a transition threshold between two operating regions. In the first region, that is, for $\lambda \in [0, \lambda_c]$, the transmissions are affected by relatively small equivalent error probabilities, $P_{eq}$. In this operating region (In what follows, this operating region will be identified by the acronym BLC, short for Below Link Capacity region.), collisions among stations occur rarely ($P_{col}$ is small) because of the reduced traffic load of the contending stations, and the stations experience good channel quality ($P_e$ is very small). For any $\lambda \leq \lambda_c$, the network is not congested, and the $N$ contending stations are able to transmit data below the link capacity limit denoted by $S_m$. This is the reason for which the aggregate throughput grows linearly with the traffic load $\lambda$.

On the other hand, the aggregate throughput tends to be upperbounded by $S_m$ in (13) for $\lambda \geq \lambda_c$. The reason is simple: the aggregate throughput in this region (In what follows, this operating region will be identified by the acronym LC, short for Link Capacity region.) is affected either by the increasing effect of the collisions among the stations, or by worse channel conditions as exemplified by $P_e \gg 0$. Further insights on this statement can be gained by the note in passing that, given $\lambda_c$, there is no need to simulate the network to obtain the aggregate throughput: it is very well approximated by the theoretical relation derived in (11) for any $\lambda \in [0, \lambda_c]$.

Once again, let us focus on the results shown in the rightmost subplots of Figure 2. The aggregate throughput gets to a maximum at a proper value of $\lambda$, above which the effect of the collisions among the stations, as well as the propagation channel, let the throughput reach a horizontal asymptote. The maximum value of $S(\lambda)$ over the whole range of values of $\lambda$ turns to be quite useful for throughput optimization. Obtained by Bianchi [2] under the hypothesis of saturated network, and considering only collisions among the stations, in our more general model such a maximum can be estimated in two steps. First, rewrite the throughput in (7) as a function of $\tau$ by using the relations that define $P_t$ and $P_e$ in terms of the probability $\tau$. Then, equate to zero the derivative of the throughput in (7) with respect to $\tau$, and obtain a solution identified by $\tau_m$. By doing so, the value of $\tau$ for which the throughput gets maximized is easily obtained:

$$\tau_m = \frac{\sigma - \sqrt{\sigma(N\sigma - 2(N - 1)(\sigma - T_t))N}}{(N - 1)(\sigma - T_t)}.$$  \hspace{1cm} (12)

Finally, evaluate the throughput in (7) on the solution $\tau_m$ noticed in (12). Upon following these steps, the maximum throughput takes on the following form

$$S(\tau_m) = S_m = \frac{E[PL]}{[T_s - (T_c/(1 - P_e)) + (T_cP_e/(1 - P_e))] + ((\sigma - T_t)(1 - \tau_m)^N + T_t)/(N\tau_m(1 - \tau_m)^{N-1}(1 - P_e))}.$$  \hspace{1cm} (13)

The critical value of $\lambda$ can be found as the abscissa where the linear model of the throughput in (9) equates to $S_m$ in (13):

$$\lambda_c = \left[ N\left( T_s - \frac{T_c}{1 - P_e} + \frac{T_cP_e}{1 - P_e} \right) \right. \left. + (\sigma - T_t)(1 - \tau_m)^N + T_t \right) \tau_m(1 - \tau_m)^{N-1}(1 - P_e)]^{-1}.$$  \hspace{1cm} (14)

The procedure for obtaining $\lambda_c$, is clearly highlighted in all the four subplots of Figure 2.

Let us investigate the behavior of the critical value $\lambda_c$ in (14) against some key network parameters.

Figure 3 shows the behavior of $\lambda_c$, as a function of the packet size $E[PL]$, for two different values of both $N$ and $P_e$ as noticed in the respective legends. Some observations are in order. $\lambda_c$ decreases for longer packet sizes $E[PL]$, as well as for increasing number of contending stations, $N$. The reason relies on the fact that for increasing packet sizes, each
whereas the horizontal lines, labeled by \( Sm \), transmitting at 1 Mbps. Other parameters are as follows: \( W_0 = 32 \), \( P_e = 0 \), and \( m = 5 \). The values of \( \lambda_c \) are noticed on the abscissa, whereas the horizontal lines, labeled by \( S_m \), represent the maximum aggregate throughput (link capacity) evaluated through (13) in each considered scenario. All the thick curves refer to \( E[PL] = 128 \) bytes, and various packet error probabilities as noticed in the figure.

Figure 4: Behavior of the aggregate throughput as a function of \( \lambda_c \) for five different values of the packet size as noticed in the legend. In each simulated scenario there are \( N = 10 \) contending stations transmitting at 1 Mbps. Other parameters are as follows: \( W_0 = 32 \), \( P_e = 0 \), and \( m = 5 \). The values of \( \lambda_c \) are noticed on the abscissa, whereas the horizontal lines, labeled by \( S_m \), represent the maximum aggregate throughput (link capacity) evaluated through (13) in each considered scenario. All the thick curves refer to \( E[PL] = 128 \) bytes, and various packet error probabilities as noticed in the figure.

Station occupies the channel longer. Such a behavior is clearly emphasized in Figure 4, where the aggregate throughput for a sample scenario comprising 10 contending stations transmitting with the packet sizes noticed in the legend, is considered. Moreover, given a payload size \( E[PL] \), \( \lambda_c \) decreases for increasing packet error probabilities, \( P_e \).

As long as the packet size increases, the aggregate throughput shows an increasing slope in the linear region characterized by traffic loads \( \lambda \in [0, \lambda_c] \), as suggested by the theoretical model in (11). Moreover, the value of \( \lambda_c \) tends to decrease because each station tends to occupy the channel longer. The three lower thick curves, labeled by \( PL = 128 \) bytes, are associated to three different values of packet error probabilities, \( P_e \). Note that the value of \( \lambda_c \) decreases for increasing values of \( P_e \), thus reducing the region characterized by small equivalent error probabilities. As long as \( P_e \to 1 \), both \( \lambda_c \) and \( S_m \) tend to zero, making the aggregate throughput vanishingly small.

Similar considerations may be derived from the behavior of \( \lambda_c \) versus \( N \) noticed in Figure 5. Roughly speaking, for fixed values of the packet size, \( \lambda_c \) tends to halve as far as the number of contending stations \( N \) doubles. The reason relies on the fact that the probability of collision increases as long as more stations try to contend for the channel.

Figure 6 compares the behaviour of \( \lambda_c \) versus \( N \) for the two different data rates 1 and 11 Mbps. From this figure we can easily note that the values of \( \lambda_c \) corresponding to the high data rate 11 Mbps are roughly three times the ones related to 1 Mbps for each number of contending stations, \( N \), and packet size \( E[PL] \) given the increased network capacity available.

4. Throughput Optimization

The considerations deduced in Section 3 are at the very basis of an optimization strategy for maximizing the aggregate throughput of the network depending on the traffic load \( \lambda \). To this end, we define two optimization strategies: Contention Window Optimization and MAC Payload Size Optimization. The first strategy is applied when the containing stations operate in the LC region, approaching the link capacity \( S_m \), whereas the second one is used for optimizing the aggregate throughput when the stations operate within the BLC region.

The next two subsections address separately the two optimization strategies, while Section 5 presents the optimization algorithm jointly implementing the two strategies.

4.1. Link Capacity Region: Contention Window Optimization.

The first optimization strategy proved to be effective for improving the aggregate throughput in the LC region, that is, for \( \lambda > \lambda_c \). The key idea here is to force the containing stations to transmit with a probability \( \tau \) equal to the one that maximizes the aggregate throughput \( S(\lambda) \). In this respect, the probability \( \tau_m \) in (12) plays a key role.
Upon considering saturated conditions, that is, imposing $q \rightarrow 1$ in (3), the probability $\tau$ can be rewritten as follows:

$$
\tau = \frac{2 \left( 1 - 2P_{eq} \right)}{(W_0 + 1) \left( 1 - 2P_{eq} \right) + W_0 P_{eq} \left( 1 - \left( 2P_{eq} \right)^m \right)}.
$$

By substituting $\tau = \tau_m$ in (4), $P_{eq}$ can be rewritten as follows:

$$
P_{eq} = 1 - (1 - \tau_m)^{N-1} + P_e - P_e \left[ 1 - (1 - \tau_m)^{N-1} \right]
= 1 + (P_e - 1)(1 - \tau_m)^{N-1} = 1 - X(P_e, \tau_m),
$$

whereby $X(P_e, \tau_m) = (1 - P_e)(1 - \tau_m)^{N-1}$.

Finally, by equating (15) to $\tau_m$ in (12), and solving for $W_0$, we obtain the optimal minimum contention window size in terms of the key network parameters:

$$
W_{OP} = \frac{1 - 2\tau_m^{-1} + X(P_e, \tau_m)(4\tau_m^{-1} - 2)}{2X(P_e, \tau_m) - 1 + (1 - X(P_e, \tau_m))(1 - 2^m(1 - X(P_e, \tau_m))^m)}.
$$

This relation yields the value of the minimum contention window $W_0$ that maximizes the aggregate throughput when the number of contending stations $N$, the packet error rate over the channel $P_e$, and the number of backoff stages $m$, are given. As a note aside, notice that, using $W_{OP}$, the maximum throughput equates to the link capacity $S_m$ in (13).

Moreover, we notice that for $m = 0$, that is, no exponential backoff is employed, $W_{OP}$ in (17) can be simplified as follows:

$$
W_{OP} = \frac{2}{\tau_m} - 1.
$$

Considering $\tau_m$ as a function of $N$, and neglecting the multiplicative constant terms, it is simple to notice that $\tau_m$ in (12) goes roughly as $1/N$ when $N \gg 1$. Therefore, the optimal contention window $W_{OP}$ grows linearly with the number of contending stations $N$ (when $N \gg 1$) in order to mitigate the effects of the collisions due to an increasing number of contending stations in the network.

In the following, we present simulation results accomplished in NS-2 for validating the theoretical models, as well as the results presented in this section. The adopted MAC layer parameters for IEEE802.11b are summarized in Table 1 [1].

The main simulation results are presented in Figures 2 and 7 in connection to the set of parameters noticed in the respective labels.
Let us focus on the results shown in Figure 7, where the saturation throughput is derived as a function of the minimum contention window size \(W_0\). The curves are parameterized with respect to three different values of \(P_e\) and \(N\) as noticed in the legend. The simulated scenario considers stations transmitting packets of size \(E[PL] = 1024\) bytes. Continuous curves represent the saturation throughput associated to the case \(P_e = 0\), the dashed curves refer to the case \(P_e = 10^{-2}\), and the dotted curves are related to the case \(P_e = 10^{-1}\).

The saturation throughput in each simulated scenario reaches a maximum corresponding to the abscissas \(W_0 = W_{OP}\) predicted by (17) and noticed beside each specific scenario in the legend. Given a predefined number of contending stations, \(N\), we notice that \(W_{OP}\) tends to be quite insensitive from the packet error rate \(P_e\), whose main effect corresponds to a reduction of the maximum achievable throughput. This behaviour holds for any given number, \(N\), of contending stations in the network. Moreover, notice that the maximum of the throughput, which settles around the value \(8.6 \times 10^5\) bps irrespective of \(N\), tends to flatten for an increasing number of contending stations, thus confirming the weak dependence of the maximum of the throughput on the value of the optimal contention window, \(W_{OP}\).

Similar considerations can be drawn from Figure 8 where the higher data rate 11 Mbps has been employed. By contrasting the curves in Figures 7 and 8, it can be easily noted that the optimal minimum contention window sizes \(W_0\) corresponding to 11 Mbps are almost 2.5 times lower than the ones achieved for the lower data rate 1 Mbps.

4.2. Below Link Capacity Region: Payload Size Optimization.

The analysis of the aggregate throughput in Section 3 revealed the basic fact that for traffic loads \(\lambda\) less than \(\lambda_c\), the network is not congested, and each station achieves a throughput roughly equal to \(E[PL] \cdot \lambda\) (the aggregate throughput is thus \(S(\lambda) = N \cdot E[PL] \cdot \lambda\)). We note in passing that the throughput does not depend on the minimum contention window \(W_0\) in the LC region. Therefore, given \(N\) contending stations, the throughput can only be improved by increasing the payload size \(E[PL]\) when the traffic load satisfies the relation \(\lambda \leq \lambda_c\).

Before proceeding any further, let us discuss two important issues in connection with the choice of the packet size \(E[PL]\).

As long as the erroneous bits are independently and identically distributed over the received packet (We notice that wireless transceivers make use of interleaving in order to break the correlation due to the frequency selectivity of the transmission channel [24].) the packet error probability \(P_e\) can be evaluated as

\[
P_e = 1 - [1 - P_e(\text{PLCP})] \cdot [1 - P_e(\text{DATA})],
\]

where,

\[
P_e(\text{PLCP}) = 1 - [1 - P_b(\text{BPSK})]^\text{PHY},
\]

\[
P_e(\text{DATA}) = 1 - (1 - P_b)^{\text{MAC} + E[PL]}.
\]

Some observations are in order. Let us focus on the results shown in Figure 2. A quick comparison among the leftmost and the rightmost subplots of Figure 2 reveals that the choice \(W_0 = W_{OP}\) in (17) guarantees improved performance for any \(\lambda > \lambda_c\), thus making the aggregate throughput equal \(S_m\). Throughput penalties due to the use of an suboptimal \(W_0\) are in the order of 100 kbps.

Notice also that the value of \(\lambda_c\) is independent from the minimum contention window chosen.

The key observation from the subplots of Figure 2 concerns the fact that the aggregate throughput of the optimized network can be modeled as follows:

\[
S(\lambda)\big|_{W_{OP}} = \begin{cases} 
N \cdot E[PL] \cdot \lambda, & \lambda \leq \lambda_c \\
S_m, & \lambda > \lambda_c.
\end{cases}
\]

As emphasized by the curves in Figure 2, this model shows a very good agreement with both theoretical and simulation results. Once again, notice that the aggregate throughput may be predicted quite accurately without resorting to simulation.
In the previous relations, MAC\(h\) and PHY\(h\) are, respectively, the sizes (in bits) of the MAC and PLCP (PLCP is short for Physical Layer Convergence Protocol) headers, and \(E[PL]\) is the size of the data payload in bits. Equation (22) accounts for the fact that a packet containing the useful data, is considered erroneous when at least one bit is erroneously received. Relation (20) is obtained by noting that a packet is received erroneously when the errors occur either in the PLCP part of the packet, or in the information data.

We notice that the relations (20) through (22) are valid either with the use of convolutional coding, whereby the bit error rate performance does not depend on the code block size [24], or for protocols that do not employ channel encoding at the physical layer.

From (22) it is quite evident that the higher the payload size, the higher the packet error rate; and vice versa. As a note aside, we notice that this behavior does not hold when concatenated convolutional channel codes [25–27], as well as low-density parity-check codes, are employed as channel codes. Indeed, for these codes the packet error rate depends on the size of the encoded block of data, with longer packets having smaller probability of error compared to shorter packets.

The second issue to be considered during the choice of \(E[PL]\) is related to the critical value \(\lambda_c\), which depends on the packet error probability (\(P_e\) defined in (20)), and consequently on the payload size. Given a traffic load \(\lambda\), any change in \(E[PL]\) could affect the network operating region, which might move from the BLC region to the LC region.

Let us discuss a simple scenario in order to reveal this issue. (Where not otherwise specified, we employ the network parameters summarized in Table 1.) Consider a network (identified in the following as scenario A) where \(N = 10\) contending stations with traffic load \(\lambda = 8\) pkt/s transmit packets of size \(E[PL] = 1024\) bytes. Assume that the bit error probability due to the channel conditions is \(P_b = 10^{-5}\), which corresponds to a packet error rate \(P_e = 8.248 \cdot 10^{-2}\) (from (20)). Upon using the network parameters summarized in Table 1, the critical load in (14) corresponds to \(\lambda_c = 9.61\) pkt/s. Since \(\lambda < \lambda_c\), the network is in the BLC operating region. If the stations increase the payload size up to \(E[PL] = 2048\) bytes, the packet error rate increases to the value \(P_e = 1.546 \cdot 10^{-1}\), with the side effect of decreasing \(\lambda_c\) to 4.71 pkt/s. Therefore, for the given traffic load \(\lambda = 8\) pkt/s, the network starts operating into the LC region.

Based on the considerations deduced in the scenario A, the proposed optimization technique aims to optimize the payload size in two consecutive steps. In the first step, we find the size of the payload in such a way that the critical threshold \(\lambda_c\) equals the actual traffic load \(\lambda\), in order for the network not to operate beyond the BLC region.

The second step verifies whether the packet error rate associated to this payload size, is below a predefined PER-target (identified by \(PER_t\)), which defines the maximum error level imposed by the application layer. First of all, given an estimated bit error probability at the physical layer, solve (22) for \(E[PL]\). Then, solve (20) for \(1 - P_e\) and substitute the relation in place of \(1 - Pr(DATA)\). Considering \(P_e = PER_t\), the following maximum payload size follows:

\[
E[PL]_{\text{max}} = \left\lfloor \frac{\ln(1 - PER_t/1 - P_e(\text{PLCP}))}{\ln(1 - P_b)} \right\rfloor - \text{MAC}_h,
\]

(23)

whereby \(\lfloor \cdot \rfloor\) is the ceil of the enclosed number.

Let us consider again the scenario A discussed previously with the application of this algorithm. Consider 10 stations transmitting packets of size \(E[PL] = 1024\) bytes at the traffic load \(\lambda = 5\) pkt/s. Assume that the bit error rate imposed by the specific channel conditions is \(P_b = 10^{-5}\), yielding \(P_e = 8.248 \cdot 10^{-2}\) and \(\lambda_c = 9.6\) pkt/s. Moreover, assume a target \(PER_t\) equal to (This is the maximum \(PER\) specified in the IEEE 802.11b standard [1], guaranteed at the receiver when the received power reaches the receiver sensitivity.) \(PER_t = 8 \cdot 10^{-2}\).

Since \(\lambda < \lambda_c\), after the first step, the algorithm selects the optimal size \(E[PL]_1 = 1938\) bytes, which moves the working point near the link capacity and leads to the actual packet error rate \(P_e = 1.47 \cdot 10^{-1}\). Since the constraint \(P_e \leq PER_t\) is not satisfied, the proposed method estimates the size \(E[PL]\) in order to attain the \(PER_t\) constraint. The packet size becomes eventually \(E[PL]_2 = 991\) bytes (from (23)), yielding a packet error rate \(P_e \approx 8 \cdot 10^{-2}\) and a \(\lambda_c^* = 9.92\) pkt/s, thus leaving the working point in the BLC region.

Finally, the optimal payload size is

\[
E[PL] = \min\left\{E[PL]_1, E[PL]_2, E[PL]_{\text{max}}\right\} = E[PL]_2 \approx 991\text{ bytes},
\]

(24)

where \(E[PL]_{\text{max}} = 2312\) bytes is the maximum packet size imposed by the standard [1].

Simulation results of a sample network employing the algorithm described above, are presented in the next section, along with a sample code fragment summarizing the key steps of the optimization technique.
In this section, we present simulation results for a network of \( N = 10 \) contending stations, employing the optimization strategies described in Section 4.

The basic steps of the proposed optimization algorithm are summarized in Algorithm 1. Let us spend a few words about the implementation of the proposed algorithm in a WLAN setting employing the infrastructure mode, where an Access Point (AP) monitors the transmissions of \( N \) contending stations.

The evaluation of the critical value \( \lambda_c \) in (14) requires the estimation of three key parameters of the network: the number of contending stations \( N \), the packet error rate \( P_e \), and the value of the optimal \( r_m \), which depends on \( N \). Each station can estimate the number of contending stations by sorting to one of the algorithms proposed in [28–30]. On the other hand, the packet error probability \( P_e \) may be evaluated through (20), once the bit error probability \( P_b \) at the physical layer is estimated. We recall that \( P_b \) can be estimated by employing proper training sequences at the physical layer of the wireless receiver [24].

Finally, the value of the optimal \( W_{\text{opt}} \) can be evaluated by each station with the parameters \( N \) and \( P_e \). A similar reasoning applies for the estimation of the optimal packet size.

We have realized a C++ simulator implementing all the basic directives of the IEEE 802.11b protocol with the 2-way handshaking mechanism, namely exponential backoff, waiting times, post-backoff, and so on. In our simulator, the optimization algorithm is dynamically executed for any specified scenario. For the sake of analyzing the effects of the optimization, the instantaneous network throughput is evaluated over the whole simulation. The aggregate throughputs obtained in the investigated scenarios are shown in Figure 9.

In the first scenario, we considered a congested network in which 10 stations transmit packet of fixed size 1028 bytes at the packet load \( \lambda = 1000 \text{ pkt/s} \) and bit rate 1 Mbps. Other parameters are \( W_0 = 32 \) and \( m = 5 \), whereas the channel conditions are assumed to be ideal. As shown in Figure 9 (curve labeled LC region with asterisk-marked points), the aggregate throughput is about \( 7.6 \times 10^5 \text{ bps} \). After 40 s, 5 out of the 10 stations turn their traffic off. Between 40 s and 80 s, the aggregate throughput increases to about \( 8.2 \times 10^5 \text{ bps} \) because of the reduced effect of the collisions among the contending stations. After 80 s, the 5 stations turn on again and the aggregate throughput decreases to \( 7.6 \times 10^5 \text{ bps} \).

Consider again the same scenario over a time interval of 120 s, whereby the contending stations adopt the optimal contention window. Upon using (17), the optimal contention window is \( W_{\text{opt}}^{0–40} = 275 \) when 10 stations transmit over the network, and \( W_{\text{opt}}^{40–80} = 130 \) during the interval in which only 5 stations contend for the channel.

The aggregate throughput in the optimized scenario is about \( 8.6 \times 10^5 \text{ bps} \), as noticed by the star-marked curve in the same figure. We notice that the optimized throughput is quite constant over the whole simulation independently on the number of contending stations in the network, approaching the theoretical maximum throughput \( S_m = 8.6 \times 10^5 \text{ bps} \) obtained from (13). Moreover, notice that the two curves \( S_m \) related to both \( N = 5 \) and 10 are almost superimposed.

Consider another scenario differing from the previous one in that the contending stations operate in the BLC region. As above, 10 stations contend for the channel in the time intervals [0,40] and [80,120] s, while during the time interval [40,80] s 5 out of the 10 stations turn off. The traffic load is \( \lambda = 8 \text{ pkt/s} \), and the packet interarrival times are exponentially distributed.

Let us focus on the aggregate throughput depicted in Figure 9. The curves related to the scenario at hand are identified by the labels \( \lambda = 8 \text{ pkt/s} \). As a result of the application of the proposed algorithm to the choice of the packet size, we obtain \( E[\text{PL}]_{\text{opt}}^{0–40} = 1383 \text{ bytes} \) for the case when 10 stations are transmitting, and \( E[\text{PL}]_{\text{opt}}^{40–80} = 2312 \text{ bytes} \) in the other scenario with 5 active stations.

![Figure 9: Aggregate throughput of a network with a maximum number of stations equal to 10, transmitting at the fixed bit rate 1 Mbps. Curve \( S_m \) represents the maximum link capacity evaluated through (13). The other network parameters are described in the paper. Curves labeled \( \text{LC region} \) refer to a scenario whereby the contending stations transmit with packet rate \( \lambda = 10^3 \text{ pkt/s} \). In all the scenarios investigated in the figure, during the time intervals [0,40] s and [80,120] s ten stations contend for the channel, while in the time interval [40,80] s only 5 stations are active. Curves labeled \( \lambda = 8 \text{ pkt/s}, \text{BLC region} \) refer to a scenario whereby the contending stations operate in the BLC region. In this situation, the optimization of the contention window does not improve the aggregate throughput and the two curves are superimposed. In the BLC region, the aggregate throughput can be improved by varying the transmission packet size as emphasized by the curve labeled \( \lambda = 8 \text{ pkt/s}, \text{BLC region}, W_{\text{opt}} \).](image)
Simulation results show that the proposed algorithm guarantees improved throughput performance on the order of 160 kbps when 10 stations are active, and about 400 kbps when only 5 stations contend for the channel. We notice in passing that, despite the optimization, the aggregate throughput could not reach the maximum $S_m$ because of the low traffic load. Indeed, the capacity link $S_m$ could be achieved by using a packet size longer than either the one imposed by the standard [1], and the value obtained from (23).

For the sake to verify that the optimization of the minimum contention window does not affect the aggregate throughput in the BLC region, Figure 9 also shows the aggregate throughput obtained by simulation using a minimum contention window equal to $W_{DP}$. Notice that the related curve is superimposed to the one related to the scenario in which the contending stations employ the minimum contention window $W_0 = 32$ suggested by the standard [1].

6. Conclusions

This paper proposed an optimization framework for maximizing the throughput of the distributed coordination function (DCF) basic access mechanism at the data link layer of IEEE 802.11 protocols. Based on the theoretical derivations, as well as on simulation results, a simple model of the optimized DCF throughput has been derived. Such a model turns to be quite useful for predicting the aggregate throughput obtained by simulation using a minimum contention window equal to $W_{DP}$. Based on the theoretical derivations, as well as on simulation results, a simple model of the optimized DCF throughput has been derived. Such a model turns to be quite useful for predicting the aggregate throughput of the DCF in a variety of network conditions.

For throughput modeling, we considered general operating conditions accounting for both nonsaturated and saturated traffic in the presence of transmission channel errors identified by the packet error rate $P_e$. Simulation results closely matched the theoretical derivations, confirming the effectiveness of both the proposed DCF model and the cross-layer optimization algorithm.

References


Submit your manuscripts at
http://www.hindawi.com