Optimizing the IEEE 802.11b Performance using Slow Congestion Window Decrease

Jérôme Galtier
France Telecom R&D
905 rue Albert Einstein
F-06921 Sophia Antipolis cedex, France
e-mail: Jerome.Galtier@francetelecom.com
Phone: +33 4 92945329, Fax: +33 4 93653591.

ABSTRACT

We investigate a new process for the decrease of the IEEE 802.11b Contention Window (CW) called CW decrease, introduced by Ni et al. We give some mathematical analysis of the model that explains the experimental results previously obtained. The analysis allows us to refine the MAC optimization and reach the optimal asymptotical saturation throughput when the number of stations increases. We also investigate other parameters such as the average waiting time, that lead us to introduce another variant of the CW decrease that we call additive - as opposed to the multiplicative original scheme. We provide some experimental results based upon simulation to validate our analysis. Surprisingly our global goodput rate in the basic mode is kept from 7.4 Mbit/s with 5 stations to more than 7.3 Mbit/s with 100 stations, overperforming by 40% the original basic scheme and by 8% the scheme with the RTS/CTS enhancement.

Keywords. Multiple access, IEEE 802.11b, MAC layer, DCF, CSMA/CA, Goodput.

1 Introduction

The 802.11b norm has received in the past years a growing interest for its ability to offer relatively wide-band radio networking. Applications cover a large area of domains including computer network wireless infrastructures, and high speed Internet access for rural areas.

The underlying mechanism of the norm[1] is a 2-layer protocol whose first part relies on a derivative of the Binary Exponential Backoff protocol (BEB). The system works as follows. Each station stores a contention window (CW) size variable, which is an integer W varying from CW_{min} to CW_{max}. Before transmitting, a station observes the channel during a silent security period called DIFS (for Distributed Inter-Frame Spacing, $50\mu s$). Then the station chooses a random number k in $\{0, \ldots, W-1\}$. The station waits during k slots (each of $20\mu s$). If the channel is occupied during one of these slots, the station stops decrementing k, waits until a new DIFS silent period is observed and then continues decrementing k if new slots are empty. Once the k empty slots have been waited, the station sends its packet. The only way to know whether the packet was received is the reception of an ACK, that is sent after a SIFS (for Short Inter-Frame Spacing, $10\mu s$) period by the receiving station. If a transmission is unsuccessful (i.e. no ACK is received and a DIFS period is observed immediately after emission), the CW size W of the sender is multiplied by a constant factor, denoted as PF (the *persistence factor*, denoted as η throughout this paper), up to CW_{max}. In the 802.11b variant of the norm, we have $PF = 2$, but other factors are considered in the 802.11e release. In all the variants, in case of success, the norm proposes to reset W to CW_{min}, which is highly aggressive to the channel, especially if it is close to congestion. Therefore an alternative approach [11] is then to set:

$$W := \max(CW_{min}, RF \cdot W),$$

where RF is an appropriate *reduction factor*. This is the basic mechanism that we study in this paper. A typical value is $RF = 1/PF$ but other alternative features will be considered. An alternative mechanism we will consider is to *add* a constant value ω to W in case of failure, and withdraw it in case of success, while staying within $[CW_{min}, CW_{max}]$.

At this point we need to say that an additional mechanism, called RTS/CTS (for *Request-To-Send/Clear-To-Send*), is included in the IEEE 802.11 families of norms. With this feature, the sender first emits a small RTS packet to warn that it will send a packet. If another station attempts to emit at the same time, noone can hear any of the messages, and the channel is free after a DIFS period. Otherwise, the receiving station sends a CTS after a SIFS period, and all the stations let the sender emit its packet till completion. Therefore, while consuming a share of the bandwidth, this mechanism allows to pay a lower price for the collisions. We need to stress that the lower the packet size is, the less efficient it is, which makes it inappropriate for real-time voice applications for instance.

Already many research work has been done to model the 802.11 window decrease process. Strong simplifying assumptions are at the basis of some models [6], while others focus on an individual station while considering that the effect of the others on the channel can be represented by an occupancy probability p (see [3, 13, 14]), following an earlier popular approach on CSMA [10, 2].

In fact, a closed loop effect naturally takes place in the 802.11 features. Coarsely speaking, when more collisions occur, the CW size in each station increases, so that the collisions are less frequent. It turns out, as derived for instance in [3], that this norm, however, is not efficient when the number of stations increases. As more and more stations access the system, the time spent to collisions and auto-regulation of the contention window increases drastically, to the point that the total amount of goodput tends to zero as the number of stations grows, resulting in a very reduced bandwidth offered to each individual.

The main reason for that is that *no memory* is kept on the *number of stations* present in the system, and as soon as a station emits one packet successfully, it forgets any information on the past and accesses again the channel with a probability that do not consider the number of stations present in the system. As a result, a number of collisions is necessary to regulate the access probability, which wastes the bandwidth for this simple purpose.

A key point in that difficulty is to know or not at any point of time the number of stations that are *active*, i.e., currently transmitting. Of course, some simple paradigm would consist to handle the problem in a centralized way such as done in [5], but then the approach looses precisely the flexibility brought by the DCF mechanism in the IEEE 802.11 family of norms (as opposed to the PCF mechanism that directs the transmissions also in a centralized basis). A series of works have addressed the problem of *estimating* the number of active stations. The seminal work can be found in [4] and [6]. In [4] the assumption is first made that all the stations are transmitting with the same CW size W, and an estimate is then done based on the number of observed occupations $c(B)$ during an observation period of B slot times. The number of stations n is then estimated by

$$n \approx 1 + \frac{E[c(B)](W+1)}{2B},$$

and in return, the size of the window W is fixed according to n. However, some instability problems oblige the authors to introduce some tricky functions that make the protocol more complex to implement and to analyze. The question becomes even more difficult when the number of active stations varies. The alternative approach of [6] consists in observing the total amount of idle slots in a transmission time and setting:

$$n \approx \frac{1-p}{Total\_Idle\_p \cdot p} = \frac{W-1}{B-c(B)},$$

using the above notation, and setting B equal to one transmission time. Here again, some smoothing factors are necessary to make the protocol stable. The authors argue that the capacity is globally preserved but the resulting protocol implementation is again quite complex.

Not only those approaches are somehow more complicated, but performance parameters become terribly difficult to analyze. For instance, it turns out that most of these protocols - if not all - seem to solve the bandwidth sharing problem by offering almost all of the capacity to one station for some (possibly long) periods of time. Some authors have defended this behavior as positive and proposed to implement it in the protocol [13]. Others view it as a negative side-effect that should be avoided [8]. In any cases, it is desirable to have a protocol that can be analyzed and tuned correctly to that respect. In particular, as will be shown in this paper, it turns out that for some versions of the protocol, a station can stay waiting for an infinite amount of time - which should be avoided.

Another more confusing aspect addresses the fairness aspects of the protocol. Indeed, as a station gets further from the others, its transmission rate decreases from 11 Mbit/s to 2 Mbit/s while the remaining stations maintain their respective rates from one to each other. As observed by [9], the actual IEEE 802.11 norm solves the problem on a Max-Min basis, that is, gives the same throughput to all the stations, regardless of the fact they can transmit to a lot more than 2 Mbit/s. We notice, by the way, that all the above approaches will have the same behavior since they do not consider the effective rate of transmission (or the transmission time of a packet) of the stations. Without addressing primarily this problem in the present paper (we will only provide individual average waiting times of the stations) we stress the importance of having a protocol that can be analyzed properly.

To that respect, the slow CW decrease approach is really attractive since the actual size of the windows keeps an implicit memory of the stations present in the system, avoiding any complicated estimation process. One can verify - either on simulations and on a mathematical analysis - that the throughput is efficiently distributed and tune the protocol to address the points mentioned above. On top of that, no more information on the system is required than the success or the failure of each individual attempt of the station to access the channel, which fits exactly to the IEEE 802.11 norm assumptions.

The paper is organized as follows. The two next sections are devoted to the mathematical analysis of the protocols, with a particular emphasis on the goodput and waiting time when the number of stations increases. The last one applies the theory into some concrete simulation scenarios - applying restricting parameters specified by the norm.

2 The multiplicative slow CW decrease model

Following [11], we represent the contention window decrease scheme by a Markov chain model where each state is indexed by two integers (i, j). The first integer i represents the stage of the process (a class of fixed contention window size W_i) and j the number of slots to wait before transmission. An additional parameter, $g = -\log(RF)/\log(PF)$ (recall $\eta = PF$), allows to regulate the decrease of the contention window size. We set $W_0 = CW_{min}$ and $W_m = CW_{max}$. We also add the correction of [14] on the fact that a station decrements its backoff counter only when no packet is heard. We use for this phenomenon the variable p_c. Note that results following the model of [11] are obtained for $p_c = 1$, while the real case is addressed by $p_c = p$. (The reader can also consider that coming from one system of formulas to the other is equivalent to increase the header of a packet by one slot of $20\mu s$.) The transition probabilities are then given as follows:

$$
\begin{array}{ll}
P[i,k|i,k+1]=1-p_c & k\in\{0,\ldots,W_i-2\}, \\
 & i\in\{0,\ldots,m\}, \\
P[i,k|i,k]=p_c & k\in\{1,\ldots,W_i-1\}, \\
 & i\in\{0,\ldots,m\}, \\
P[0,k|i,0]=(1-p)/W_0 & k\in\{0,\ldots,W_0-1\}, \\
 & i\in\{0,\ldots,g-1\}, \\
P[i-g,k|i,0]=(1-p)/W_{i-g} & k\in\{0,\ldots,W_{i-g}-1\}, \\
 & i\in\{g,\ldots,m\}, \\
P[i,k|i-1,0]=p/W_i & k\in\{0,\ldots,W_i-1\}, \\
 & i\in\{1,\ldots,m\}, \\
P[m,k|m,0]=p/W_m & k\in\{0,\ldots,W_m-1\}.
\end{array}
$$

We plot this process in Fig. 1 in the case $g = 1$ - which will be the most important one studied in this paper. Let $\{s(t), b(t)\}$ be a bi-dimensional stochastic process that follows these laws. Markovian properties show that, in this case, the process converges almost surely to a stationary distribution, given by

$$\pi_{i,k} = \lim_{t\to\infty} P[s(t) = i, b(t) = k], \quad \substack{i\in\{0,\ldots,m\} \\ k\in\{0,\ldots,W_i-1\}}.$$

The analysis allows to derive the following formulas:

$$
\begin{array}{rcl}
\pi_{0,0} & = & (1-p)\sum_{j=0}^{g}\pi_{j,0}, \\
\pi_{i,0} & = & p\pi_{i-1,0} + (1-p)\pi_{i+g,0} \\
 & & \text{for } 0 < i \le m-g, \\
\pi_{i,0} & = & p\pi_{i-1,0} \\
 & & \text{for } m-g < i < m, \\
p\pi_{m-1,0} & = & (1-p)\pi_{m,0}.
\end{array}
$$

2.1 Closed expressions

We derive some close expressions for small values of g. If we set $g = 1$, then we have:

$$
\begin{array}{l}
\pi_{m-1,0} = \frac{1-p}{p}\pi_{m,0} \\
\pi_{i,0} = p\pi_{i-1,0} + (1-p)\pi_{i+1,0}, \; 0 < i < m.
\end{array} \tag{1}
$$

So we write $\pi_{m-j,0} = ar_1^j + br_2^j$, for $a, b \in \mathbb{R}$, where r_1 and r_2 are the roots of the polynomial

$$pX^2 - X + (1-p).$$

We then derive:

Result 1 *For $g = 1$, the distribution $\pi_{\cdot,\cdot}$ follows the equation :*

$$\pi_{m-j,0} = \left(\frac{1-p}{p}\right)^j \pi_{m,0}, \quad 0 \le j \le m. \tag{2}$$

It is then not difficult to see that, for all $i \in \{0, \ldots, m\}$,

$$\sum_{k=0}^{k=W_i-1} \pi_{i,k} = \left(1 + \frac{W_i - 1}{2(1-p_c)}\right)\pi_{i,0}.$$

and from that value we can derive the value of $\pi_{m,0}$, given that the congestion window size is given by $W_i = \eta^i W_0$ for some number $\eta > 1$.

Result 2 *For $g = 1$, we have:*

$$\pi_{m,0} = \frac{2\cdot p^m(1-p_c)}{\frac{1-2p_c}{1-2p}((1-p)^{m+1}-p^{m+1}) + W_0\frac{(1-p)^{m+1}-(\eta p)^{m+1}}{1-(\eta+1)p}}. \tag{3}$$

A closed expression for the probability τ that the terminal attempts to send a packet can be derived:

Result 3 *For $g = 1$, the probability that a station transmits in a randomly chosen slot time is given by:*

$$\tau = \frac{2(1-p_c)}{1-2p_c + W_0 \frac{1-2p}{1-(\eta+1)p} \frac{(1-p)^{m+1}-(\eta p)^{m+1}}{(1-p)^{m+1}-p^{m+1}}}. \quad (4)$$

PROOF. We simply derive the formula from $\tau = \sum_{j=0}^{m} \pi_{j,0}$. □

A similar analysis can be done for all the values of g. The closed expression can be expected till $g = 4$, after what several mathematical difficulties can occur, as stated by the Galois theorem. Anyway, the analysis is still valid after that value via a numerical extraction of the roots of a polynomial.

For instance, for $g = 2$, the characteristic polynomial is given by:

$$pX^3 - X^2 + (1-p),$$

which leads to:

Result 4 *For $g = 2$, the distribution $\pi_{\cdot,\cdot}$ is given by :*

$$\pi_{m-j,0}/\pi_{m,0} = \left[\frac{1}{2}\left(1+\sqrt{\frac{1-p}{3p+1}}\right)\left(\frac{1-p+\sqrt{(1+3p)(1-p)}}{2p}\right)^j + \frac{1}{2}\left(1+\sqrt{\frac{1-p}{3p+1}}\right)\left(\frac{1-p-\sqrt{(1+3p)(1-p)}}{2p}\right)^j\right]. \quad (5)$$

2.2 Asymptotical behavior

In the remaining of this paper, we will study more precisely the case $g = 1$. As our main interest in the slow decrease process is the memory aspects of it, we can consider the case $m \to \infty$. The asymptotical function can be rewritten as:

$$\tau = \frac{2(1-p_c)}{1-2p_c + W_0 \frac{1-2p}{1-(\eta+1)p}} \text{ for } p < \frac{1}{\eta+1}, \quad (6)$$

and 0 otherwise. We plot the τ function in Fig. 2. We note that the value for $p = 0$ does not depend on η, but the function abruptly reaches 0 when $p = 1/(\eta+1)$. Following the analysis of [3], let n the number of active stations in the channel, we seek for the intersection point of one of (4) or (6) with the function giving the probability that the channel is free, given that all the remaining $n-1$ stations have the same probability of access τ, and all the accesses are independent processes (which is obviously an approximation):

$$p = 1-(1-\tau)^{n-1} \quad (7)$$

or, with respect to τ

$$\tau = 1-(1-p)^{\frac{1}{n-1}}. \quad (8)$$

Result 5 *If p_c is a non-decreasing function of p, the functions (8) and (4) have a unique intersection point.*

We note, in particular, that both the cases $p_c = p$ and $p_c = 1$ are handled by the theorem above.
PROOF. Obviously the function (8) is continuous, strictly increasing on $[0, 1]$ from 0 to 1. The function (4) can be rewritten as follows:

$$\tau = \frac{1}{1 + \frac{1}{2(1-p_c)}\left(W_0 \frac{1+\cdots+(\frac{(1-p)}{p\eta})^i+\cdots+(\frac{(1-p)}{p\eta})^m}{1+\cdots+(\frac{(1-p)}{p})^i+\cdots+(\frac{(1-p)}{p})^m} - 1\right)}.$$

In order to show that the intersection point exists and is unique, it sufficient to show that τ is non-increasing. Since p_c is non-decreasing with p and $W_0 \geq 1$, it is sufficient to show that the function $x \mapsto \frac{1+\cdots+(\eta x)^i+\cdots+(\eta x)^m}{1+\cdots+x^i+\cdots+x^m}$ is greater than 1 and non-decreasing as long as $\eta \geq 1$. Then use the following notations:

$$f_k(x) = \sum_{j=k+1}^{j=m} x^j a^k \qquad g_k(x) = \sum_{i=0}^{i=k} x^i a^i$$

and noticing that $g_k(x) + af_k(x) = g_{k+1}(x) + f_{k+1}(x)$, we have

$$\frac{1+\cdots+(\eta x)^i+\cdots+(\eta x)^m}{1+\cdots+x^i+\cdots+x^m} = \prod_{k=0}^{m-1} \frac{g_k(x)+af_k(x)}{g_k(x)+f_k(x)}.$$

It is then sufficient to show that each member of the right-hand product is greater than 1 and increasing. We can easily see it is greater than 1. We note also that the derivate has the sign of $(\eta-1)(f_k'(x)g_k(x) - g_k'(x)f_k(x))$. Recall that $\eta \geq 1$ and

$$\begin{aligned} & f_k'(x)g_k(x) - g_k'(x)f_k(x)) \\ & \quad = \sum_{j=k+1}^{j=m} jx^{j-1}a^{k+1} + \sum_{i=1}^{i=k}\sum_{j=k+1}^{j=m} (j-i)a^{k+1}x^{i+j-1} \\ & \quad \geq 0, \end{aligned}$$

and therefore the function is non-decreasing. □

We leave to the reader the proof of the following result:

Result 6 *The functions (8) and (6) have a unique intersection point.*

From that on, we characterize the intersection point (p^*, τ^*) when the number n of stations becomes large. From equation (6) we see that the intersection point p^* will verify $p < 1/(1+\eta)$. Substituting in (8) we have

$$\tau^* < 1 - \left(\frac{\eta}{\eta+1}\right)^{\frac{1}{n-1}} \to 0,$$

when n becomes large. As a result,

Result 7 *Under our model's assumptions (i.e. supposing that the stations' processes to access the channel are independent), and with an infinite maximum window size, the convergence point verifies* $p^* \to 1/(1+\eta)$ *when* n *becomes large, and*

$$\tau^* = \frac{1}{n}\ln\left(\frac{\eta+1}{\eta}\right) + o\left(\frac{1}{n}\right). \tag{9}$$

This result is extremely important from the theoretical point of view. It shows that naturally the process tends to share the bandwidth between the processors. We show in the next section that this indeed governs the saturation throughput when the number of stations becomes large.

2.3 Saturation throughput

Following the analysis of many authors, such as [10, 3], we can study a *renewal process* drawn in Fig. 3, that can be described as follows:

the W state is when the system has an idle slot (wait),

the Q state is when the system has had at least one query, but it is not known if a conflict has occurred (query),

the P state is when the system successfully transmits a packet (payload).

Theoretically speaking, it is a semi-Markov process, and the (discrete) transition probabilities are given by the arrows. Noting π_w, π_q and π_p the stationary probabilities of the corresponding discrete Markov chain ($\pi_w + \pi_q + \pi_p = 1$), we have the following transition equations:

$$\begin{aligned}
\pi_w &= \pi_p + (1-\tau)^n \pi_w + \left[1 - \tfrac{n\tau(1-\tau)^{n-1}}{1-(1-\tau)^n}\right]\pi_p, \\
\pi_q &= \left[1-(1-\tau)^n\right]\pi_w, \\
\pi_p &= \left[\tfrac{n\tau(1-\tau)^{n-1}}{1-(1-\tau)^n}\right]\pi_q.
\end{aligned} \tag{10}$$

Therefore, if we note μ_w, μ_q and μ_p the expected time spent in a state before a transition during the process, the proportion of time spent in the stage P (noted Thr as it is the effective saturation throughput of the channel) will be given by (for more details see for instance [12, pp. 130-132]):

$$Thr = \frac{\mu_p \pi_p}{\mu_w \pi_w + \mu_q \pi_q + \mu_p \pi_p} \tag{11}$$

Solving (10) and substituting in (11) we obtain:

$$Thr = \frac{1}{1 + \frac{\mu_w + \mu_q - \mu_q(1-\tau)^n}{\mu_p n\tau(1-\tau)^{n-1}}} \tag{12}$$

The optimal τ^+ for the throughput is the one that minimizes the denominator and after some manipulation, we can see that if it exists it verifies:

$$\mu_w + \mu_q - \mu_q(1-\tau^+)^n - n\tau^+(\mu_w + \mu_q) = 0 \tag{13}$$

From that we can derive that

$$\tau^+ = \frac{1}{n}\left(1 - \frac{\mu_q}{\mu_w+\mu_q}(1-\tau^+)^n\right) \le \frac{1}{n},$$

and $\tau^+ = \frac{K}{n} + o(\frac{1}{n})$ for some $K \ge 0$ and the K constant verifies:

$$(1-K)e^K = \frac{\mu_q}{\mu_w + \mu_q}.$$

The solution can be found quite easily by introducing the Lambert's w function (or the Omega function) satisfying $w(z)e^{w(z)} = z$ and that can be developed as

$$w(x) = \sum_{p\ge 1} \frac{(-p)^{p-1}}{p!} x^p.$$

Then the optimal τ^+ can be approximated by

$$\tau^+ = \frac{1 + w\left(-\frac{1}{e}\frac{\mu_q}{\mu_w+\mu_q}\right)}{n} + o(\frac{1}{n}) \tag{14}$$

(where $e = exp(1)$) and the corresponding saturation throughput tends to

$$Thr^+ = \left(1 - \frac{\mu_q/\mu_p}{w(-\frac{1}{e}\frac{\mu_q}{\mu_w+\mu_q})}\right)^{-1}. \tag{15}$$

Making the connection between the equations (9) and (14), we obtain:

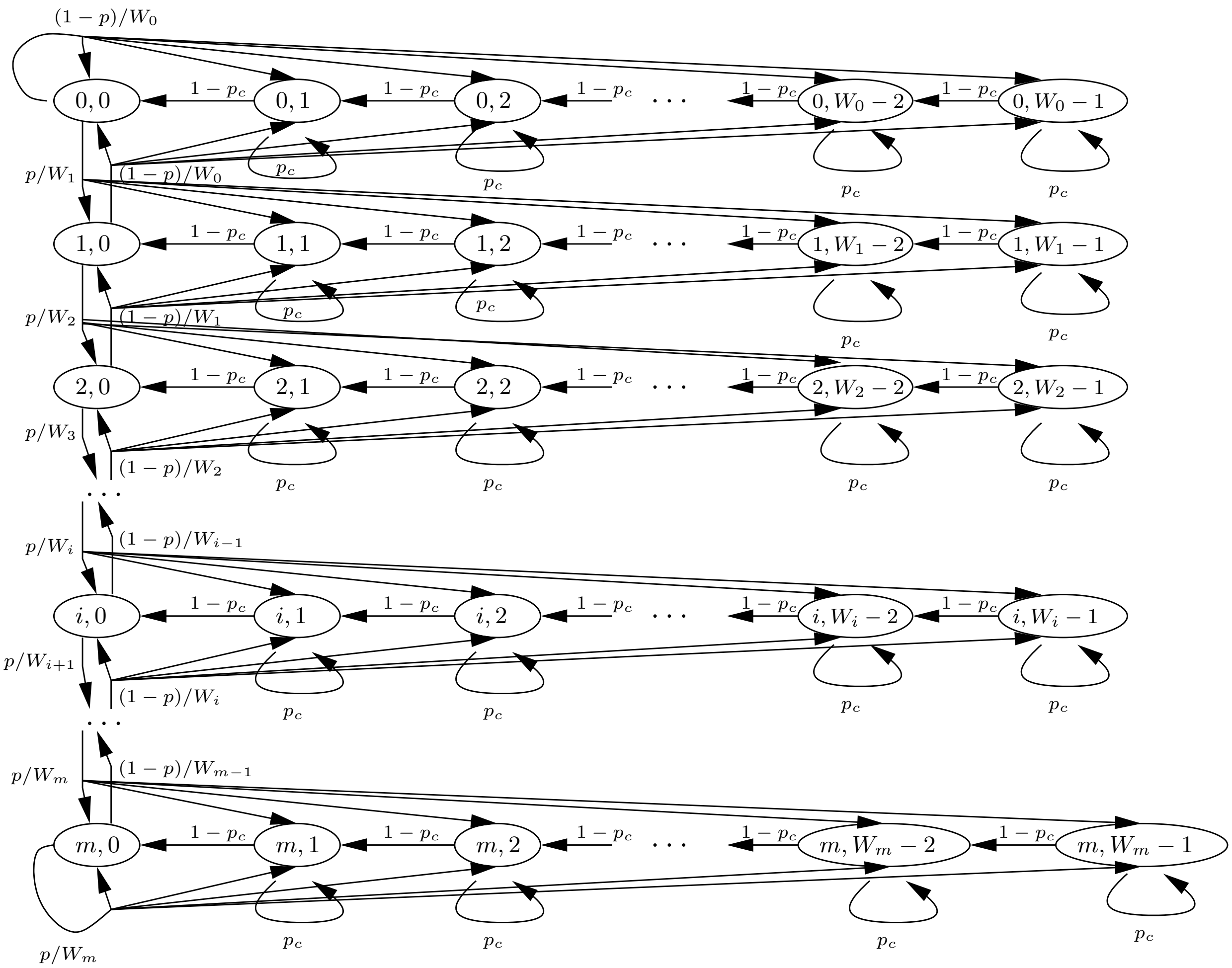

Figure 1: The Markov chain in the case $g = 1$.

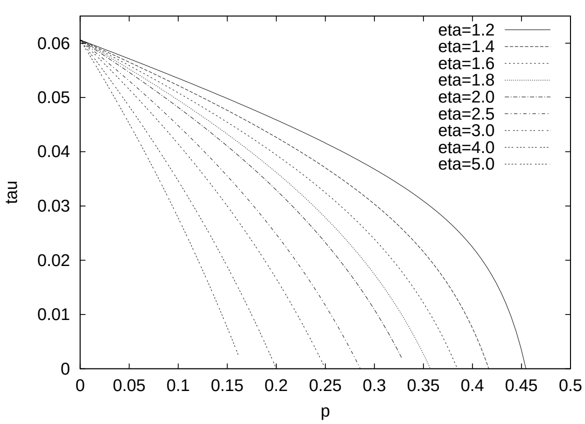

Figure 2: Different values of τ ($W_0 = 32$, $p_c = p$).

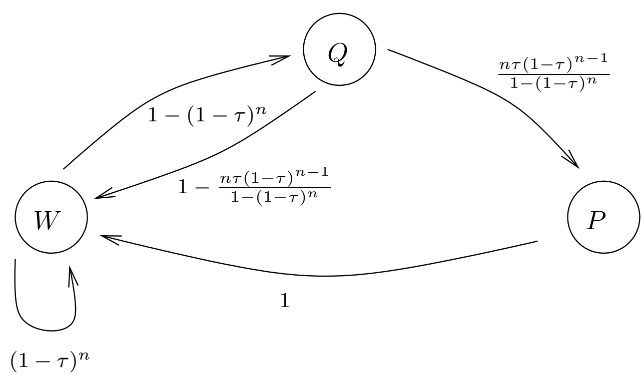

Figure 3: General figure of the slotted CSMA/CA process.

Result 8 *Under our model's assumptions (i.e. supposing that the stations' processes to access the channel are independent), the optimal saturation throughput of the channel will be reached automatically in accordance with the number of stations by setting*

$$\eta = \left(\frac{-\frac{\mu_q}{\mu_w+\mu_q}}{w\left(-\frac{1}{e}\frac{\mu_q}{\mu_w+\mu_q}\right)} - 1 \right)^{-1} \tag{16}$$

along with $m = \infty$.

At first glance, one can be surprised that W_0 plays no role to reach the asymptotical saturation throughput. This fact is not so strange because as the number of stations increases, the process rebalances the weights in higher states than W_0 and reaches this equilibrium. In Fig. 5, we plot the function $x \mapsto 1/(-x/w(-x/e) - 1)$ that governs the value of η in function of $\mu_q/(\mu_w + \mu_q)$. We see that the condition $\eta > 1$ turns to translate to $\mu_q/(\mu_w + \mu_q) > 0.613\ldots$, a condition that is easily full-filed, since the query time implies that an acknowledgment is sent, and therefore $\mu_q > 2\mu_w$.

Anyway, we show in the next paragraph that this mechanism has nevertheless some significant drawbacks that make its use significantly sensitive.

2.4 Asymptotical waiting time

An important parameter to study is the asymptotical waiting time when the number of station increases. Naturally, it is clear that the more stations are present in the system, the less probability there is that a station will access the system. We show, however, that the asymptotical behavior of the present mechanism is much worse.

Result 9 *Suppose a station follows a slow CW decrease mechanism with $m = \infty$, and observes a collision probability of $p \geq 1/(1+\eta^2)$. Then the average number of waited timeslots before emission is infinite. If $p < 1/(1+\eta^2)$, the average number of waited timeslots is given by:*

$$E[W] = \frac{1}{6(1-p_c)} \frac{\frac{1}{1-(\eta^2+1)p} - \frac{1}{1-2p}}{\frac{1-2p_c}{1-2p} + W_0 \frac{1-p}{1-(\eta+1)p}}. \tag{17}$$

Else if $m < \infty$, then the number of waited timeslots is given by:

$$E[W] = \frac{1}{6} \frac{\frac{1}{1-p_c}\left(\frac{(1-p)^{m+1}-(\eta^2 p)^{m+1}}{1-(\eta^2+1)p} - \frac{(1-p)^{m+1}-p^{m+1}}{1-2p}\right)}{\frac{1-2p_c}{1-2p}((1-p)^{m+1}-p^{m+1}) + W_0 \frac{(1-p)^{m+1}-(\eta p)^{m+1}}{1-(\eta+1)p}}. \tag{18}$$

Equation (17) shows that the waiting time tends to infinity when $p \to 1/(1+\eta^2)$. As a result, when the number of stations increases, p grows from 0 to $1/(1+\eta)$, leaving $E[W]$ infinite. This result clearly shows a major drawback of the slow CW decrease mechanism, since there is then no guarantee that a packet will be emitted within a short amount of time. The surprising point with this result is that since $1/(\eta^2+1) > 1/(\eta+1)$, a finite number of stations is sufficient to reach this behavior.

Proof. The expected waited time before leaving a state (i,k), $k \geq 0$, is given by

$$\sum_{i=0}^{\infty} (i+1) p_c^i (1-p_c) = 1/(1-p_c),$$

and therefore the general expected waited time is

$$E[W] = \sum_{i=0}^{i=m} \sum_{k=0}^{k=W_i-1} k\pi_{i,k}/(1-p_c). \tag{19}$$

Given that for $k \geq 1$, $\pi_{i,k} = \pi_{i,0}(W_i - k)/W_i/(1-p_c)$ and

$$\sum_{k=1}^{k=i-1} k^2 = \frac{2i^3 - 3i^2 + i}{6},$$

and associating to equation (2) we obtain:

$$E[W] = \sum_{i=0}^{i=m} \frac{\pi_{0,0}}{(1-p_c)^2} \left(\frac{p}{1-p}\right)^i \left[\frac{W_i^2 - 1}{6}\right]. \tag{20}$$

We set $W_i = \eta^i W_0$. We note that for $p \in (1/(\eta^2+1); 1/(\eta+1))$, the series does not converge when $m \to \infty$. Also, reintroducing (3), we obtain (18) after some manipulation. For $p < 1/(\eta^2+1)$, since the series is bounded in absolute value, letting $m \to \infty$ gives (17). Hence the result. □

This last result advocates for a mechanism that will respect waiting times more fairly. In the next section we present an alternative mechanism that keeps the good properties of the slow decrease paradigm while avoiding the presented drawbacks.

3 An additive increase/decrease model

In this section, we simply set $W_i = W_0 + \omega \cdot i$ (instead of $W_i = \eta^i W_0$) and show the behavior obtained by this assumption. We take here the opposite course to some fashion that was introduced by the MIMD mechanisms that work well indeed under some specific conditions [7]. We argue that, however, in our case the conditions are somehow different. Indeed the maximum desirable CW size is a critical parameter since a too high value will induce huge delays for some of the terminals, and a restricted one will erase the memory effects of the mechanism, which is what we try to stimulate in the present study. In that context, lowering the access to high values of CW seems to be reasonable.

3.1 The model studied

We slightly change the previous model as follows. We replace the previous parameter p by two parameters p and q, and study the following variant of the process. When the process is in a state $(i, 0)$, it goes with probability p to the upper stage, with the probability q on the lower stage, and stays in the same stage with probability $(1 - p - q)$. The evolution equations can be rewritten as follows (with $g = 1$):

$$\begin{array}{ll} P[i,k|i,k] = p_c & {\scriptstyle k\in\{1,\ldots,W_i-1\},} \\ & {\scriptstyle i\in\{0,\ldots,m\},} \\ P[i,k|i,k+1] = 1-p_c & {\scriptstyle k\in\{0,\ldots,W_i-2\},} \\ & {\scriptstyle i\in\{0,\ldots,m\},} \\ P[0,k|0,0] = (1-p)/W_0 & {\scriptstyle k\in\{0,\ldots,W_0-1\},} \\ P[i-1,k|i,0] = q/W_{i-1} & {\scriptstyle k\in\{0,\ldots,W_{i-1}-1\},} \\ P[i,k|i-1,0] = p/W_i & {\scriptstyle k\in\{0,\ldots,W_i-1\},} \\ & {\scriptstyle i\in\{1,\ldots,m\},} \\ P[i,k|i,0] = (1-p-q)/W_i & {\scriptstyle k\in\{0,\ldots,W_i-1\},} \\ & {\scriptstyle i\in\{1,\ldots,m-1\},} \\ P[m,k|m,0] = (1-q)/W_m & {\scriptstyle k\in\{0,\ldots,W_m-1\}.} \end{array} \tag{21}$$

We plot this process in Fig. 4. Let $\{s(t), b(t)\}$ be a bi-dimensional stochastic process that follows these laws. Again, markovian properties show that the process converges almost surely to a stationary distribution, given by

$$\pi_{i,k} = \lim_{t\to\infty} P[s(t) = i, b(t) = k], \quad {\scriptstyle i\in\{0,\ldots,m\}} \quad {\scriptstyle k\in\{0,\ldots,W_i-1\}}.$$

And similar formulas can be derived:

$$\begin{array}{rcl} p\pi_{0,0} &=& q\pi_{1,0}, \\ (p+q)\pi_{i,0} &=& p\pi_{i-1,0} + q\pi_{i+1,0} \\ && \text{for } 0 < i < m, \\ q\pi_{m,0} &=& p\pi_{m-1,0}. \end{array} \tag{22}$$

We then derive:

Result 10 *The distribution $\pi_{\cdot,\cdot}$ follows the equation :*

$$\pi_{j,0} = \left(\frac{p}{q}\right)^j \pi_{0,0}, \quad 0 \le j \le m. \tag{23}$$

Then the basic state is obtained as:

Result 11 *We have:*

$$\begin{array}{rcl} \frac{1}{\pi_{m,0}} &=& \frac{\left[\left(1+\frac{W_0-1}{2(1-p_c)}\right)(q-p)+\frac{pw}{2(1-p_c)}\right]q^{m+1}}{p^m(q-p)^2} \\ && -\frac{\left[\left(1+\frac{W_0+m\omega-1}{2(1-p_c)}\right)(q-p)+\frac{q\omega}{2(1-p_c)}\right]p^{m+1}}{p^m(q-p)^2}. \end{array} \tag{24}$$

Proof. Similarly as before, we have

$$1 = \sum_{i=0}^{i=m} \left(1 + \frac{W_0 + i\omega - 1}{2(1-p_c)}\right)\left(\frac{q}{p}\right)^{m-i} \pi_{m,0}.$$

And the result comes from the simple fact (that the reader can check by induction) that:

$$\sum_{i=1}^{i=m} ix^i = \frac{x - (m+1)x^{m+1} + mx^{m+2}}{(1-x)^2}.$$

□

And consequently we have the similar result:

Result 12 *If we denote as before τ the probability that a terminal will access the channel, given that it observes a probability p of failure, we have:*

$$\tau = \frac{1}{1 + \frac{W_0-1}{2(1-p_c)} + \frac{\omega}{2(1-p_c)}\frac{p}{q-p}\frac{1-(m+1)\left(\frac{p}{q}\right)^m + m\left(\frac{p}{q}\right)^{m+1}}{1-\left(\frac{p}{q}\right)^{m+1}}}. \tag{25}$$

3.2 Optimal saturation throughput

If then we tend to an asymptotical result, i. e. $m = \infty$, we have for $p < q$:

$$\tau = \frac{1}{1 + \frac{W_0-1}{2(1-p_c)} + \frac{\omega}{2(1-p_c)}\frac{p}{q-p}}. \tag{26}$$

In Fig. 6 we draw various curves for τ depending on ω. As expected, τ tends to zero as long as $p = q$. The exact value of ω is a critical parameter for the dynamicity of the process. A too small value will lead to a slow convergence of the rates (and therefore many additional collisions before regulating the value). A too high one will generate for at least some of the processes excessively high levels of CW, and therefore high levels of delay.

The idea for the use of q is to turn it into an action in case of success. To that purpose, we can write $q = (1-p)(1-\delta)$, where $\delta \in [0; 1]$ is a parameter that drives between the action of decreasing the window or leaving it at the same size. The condition $p = q$ turns to become $p = (1 - \delta)/(2 - \delta)$. Given that, we look for an asymptotical intersection point

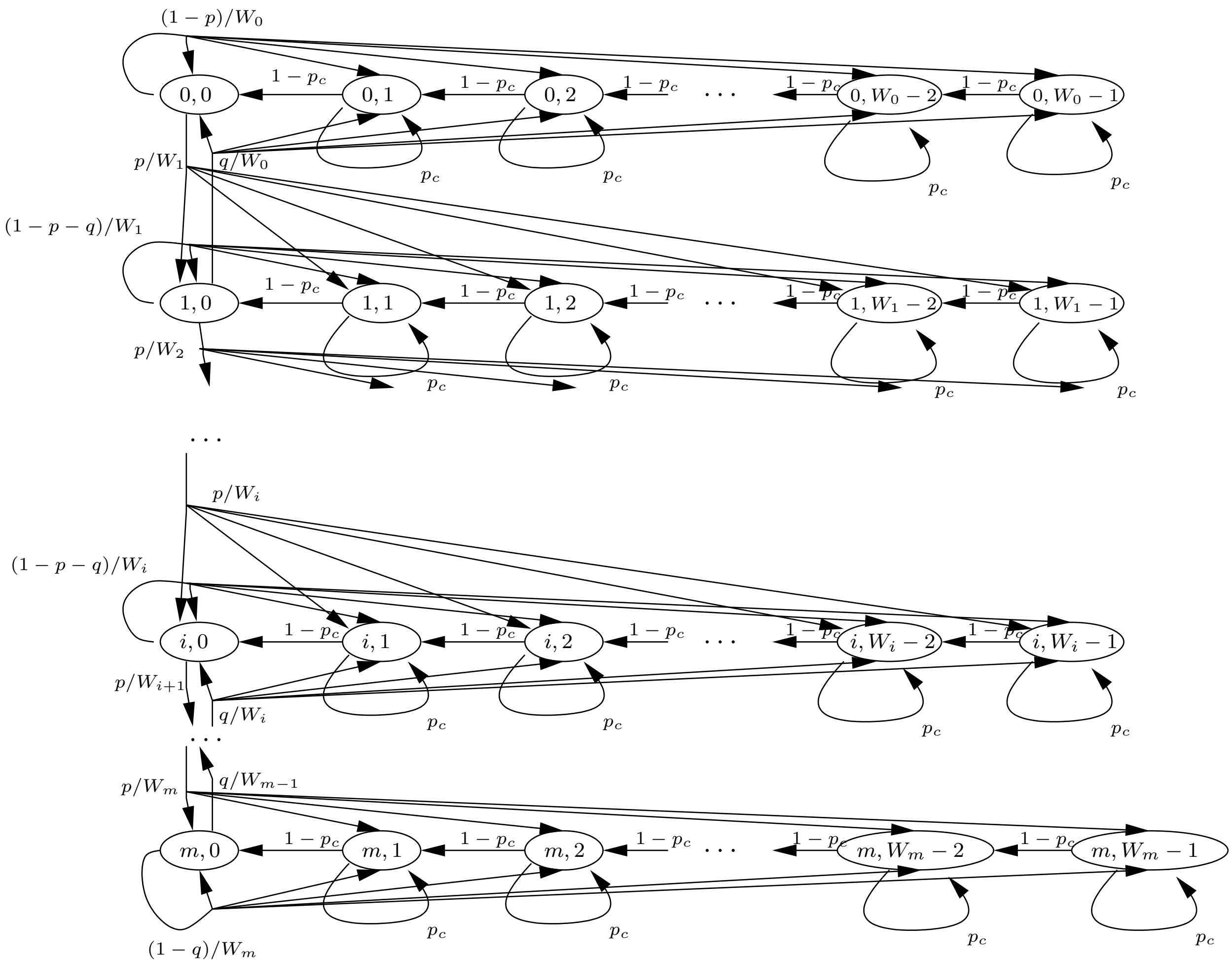

Figure 4: The Markov chain for the additive model.

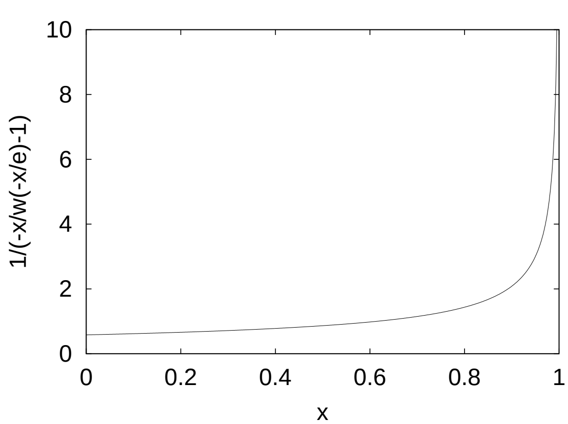

Figure 5: Optimal curve for η in function of $\mu_q/(\mu_w+\mu_q)$.

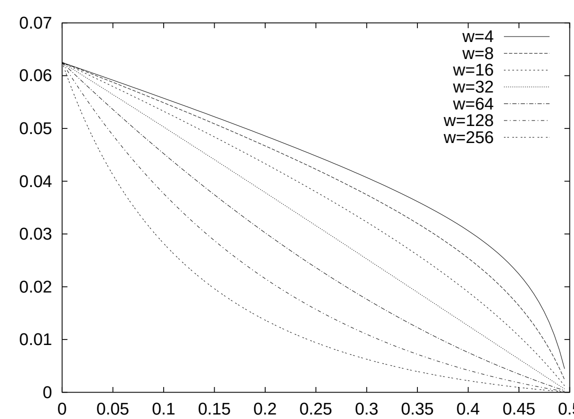

Figure 6: Different access probabilities for $W_0 = 32$, $p_c = p$, and $q = 1 - p$.

in (p, τ) with equation (8), where n is the number of terminals. Similarly as before, when n becomes large, we obtain $p \to (1-\delta)/(2-\delta)$ and

$$\tau = \frac{1}{n}\ln(2-\delta) + o\left(\frac{1}{n}\right). \quad (27)$$

At that point we see that the value of q directs the access probability to the channel. Then the optimal mode can be reached asymptotically as explained in the following.

Result 13 *Under an additive model to add/decrease the congestion window, and supposing the stations' processes to access the channel are independent, then the optimal saturation throughput can be obtained by flipping a coin that will transform an increase action into a stationary action with probability:*

$$\delta = 2 + \frac{\frac{\mu_q}{\mu_w+\mu_q}}{w\left(-\frac{1}{e}\frac{\mu_q}{\mu_w+\mu_q}\right)} \quad (28)$$

The corresponding saturation throughput is again given by the equation (15).

3.3 Asymptotical waiting time

We here give some considerations on the waiting time that prove the correctness of the additive slow CW decrease mechanism. In order to simplify the notations, we switch back to the notations with q and give some analytical evidence on the properties in question.

Result 14 *A terminal following an additive slow CW decrease Markov process observing a probability $p < q$ of collision and with an infinite number m of states experiences an average waiting number of slots of*

$$E[W] = \frac{q(q-p)(W_0^2+1) + 2W_0\omega pq + \frac{\omega^2 pq(3p-q)}{q-p}}{3(1-p_c)^2\left[(q-p)\left(2+\frac{W_0-1}{1-p_c}\right)+\frac{\omega p}{1-p_c}\right]}. \quad (29)$$

As a result, when $p \to \frac{1}{2}$, we have

$$E[W] \simeq \frac{\omega}{3}\frac{1}{q-p}. \quad (30)$$

This result is really important since it shows that asymptotically, the waiting times behaves like the inverse of $q-p$, which grows like the number of stations. Therefore, the bandwidth is not only efficiently shared, but each of the stations will receive on average a even part of it and will observe the same average waiting time.

Proof. (short) Summing up the waited slots by stages of the Markov process, we obtain similarly as before:

$$E[W] = \frac{\pi_{0,0}}{(1-p_c)^2}\sum_{i=0}^{i=m}\left(\frac{p}{q}\right)^i \frac{W_i^2-1}{6} \quad (31)$$

The essential ingredients are manipulations similar to those previously done. A useful formula that is necessary for the proof is simply:

$$\sum_{i=1}^{i=m} i^2 x^i = \frac{-x+3x^2+(m+1)^2x^{m+1}}{(1-x)^3} - \frac{(2m^2+6m+3)x^{m+2}+m(m+4)x^{m+3}}{(1-x)^3}.$$

□

In the next section we give some experimental evidence on the accuracy of these models.

4 A simulated slow CW decrease mechanism compared to the analysis

We have implemented a small simulator of the 802.11b mechanism, that takes into account all the details of the norm. The channel bit rate is set to 11 Mbit/s, the packet payload to 1500 bytes, short headers are used, and the propagation delay is neglected. Each station is managed individually and accesses to the channel in a slotted fashion with respect to its own counter. We compare in the test three algorithms:

The basic 801.11b scheme with $PF = 2$, $RF = 0$, $CW_{min} = 32$, and $CW_{max} = 1024$. In case of failure of transmission in state CW_{max}, we choose to stay in this stage.

A slow CW decrease multiplicative mechanism with $\eta = PF = 1/RF$ (i.e. $g = 1$), $CW_{min} = 32$, and $CW_{max} = 1024$. According to the previous analysis, and based upon a rate of 11 Mbit/s and an average packet size of 1500 bytes, we set $\eta = 5.5$ on the basic mode, and $\eta = 2$ with RTS/CTS.

A slow CW decrease additive mechanism with $\omega = 32$, $CW_{min} = 32$, and $CW_{max} = 1024$. According to the previous analysis, and based upon a rate of 11 Mbit/s and an average packet size of 1500 bytes, we set $\delta = 0.81910$ on the basic mode, and $\delta = 0.49434$ with RTS/CTS. In case of success, with probability δ, we do not change the CW size, and with probability $1-\delta$, we decrease it by ω. In case of failure, CW is increased by ω.

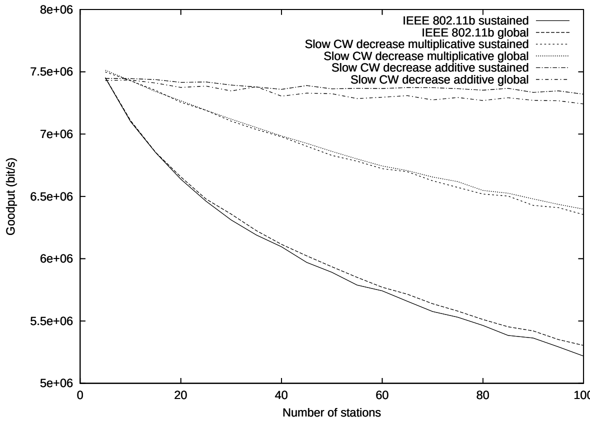

Figure 7: Simulation of the obtained rates with the basic mechanism.

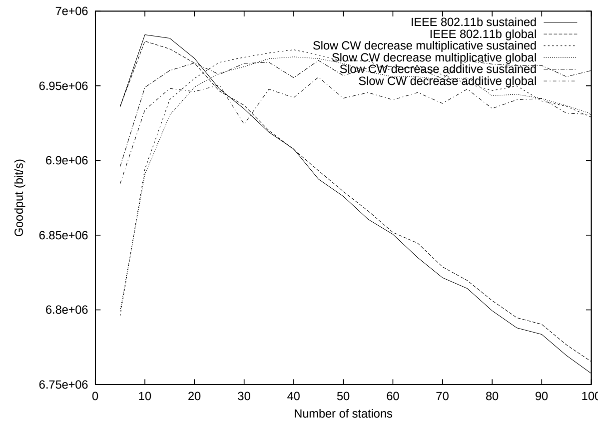

Figure 8: Simulation of the obtained rates with the RTS/CTS mechanism.

Given the relatively small value of CW_{max}, it is not realistic to present experiments with more than 100 stations. However, further experiments with increased CW_{max} showed that our approach is also satisfactory for several hundreds of stations.

In the first experiment, we give to each of the n stations a volume of 100 Mbytes/n to transmit. We let them the simultaneous access to the channel, and we compute the total volume of goodput divided by the time necessary to complete all the transmissions - we call that number the *global rate*. We also compute the total number of packets that were sent during a given period after stabilization of the processes, that is between $t = 50s$ and $t = 100s$ - we call the obtained rate the *sustained rate*. The difference between these two parameters indicates the speed of convergence of the process. We plot the results in Figs. 7 and 8. For the basic mechanism (Fig. 7), the gains obtained with the additive mechanism are of more than 40% for 100 stations compared to the basic mechanism, and 3% with the RTS/CTS mechanism (while losing 1% for small number of stations). We note that the obtained rate without RTS/CTS is on top of 8% of the norm using RTS/CTS. This shows that even if RTS/CTS cuts the optimization, the improvement is such that the basic mechanism without RTS/CTS becomes more efficient. In this context, only the hidden terminal problem justifies the use of RTS/CTS.

In the second experiment, we start with a bench of 100 stations having each 1 Mbyte to transmit. Then at time $t = 50s$ a new series of 10 stations (numbered from 100 to 109) arrive with a volume of 100 kbytes to transmit. We plot the time of completion of the processes for the different stations in Figs. 9 and 10. We observe - generally speaking - that the IEEE 802.11 standard and the slow CW decrease multiplicative mechanism experience a larger deviation than that of the additive mechanism. However, we notice that the slow CW decrease additive mechanism observes a larger deviation for the first 100 stations with the RTS/CTS mechanism. This and Fig. 8 showing some larger difference between the global and the sustained rates give some insight that the process does not converge very quickly to the asymptotical behavior. This is probably due to the introduction of the q parameter that lowers the impact of success and failure rates in the congestion window evolution. It suggests that if this problem becomes critical, more information on the channel should be integrated in the process.

5 Conclusion

In this abstract, we have derived exact formulas for small values of g for the slow congestion window decrease 802.11 WLAN mechanism, with multiplicative and additive evolution mechanisms. Our analysis has shown that both can naturally reach the asymptotical optimal saturation throughput of the channel regardless of the number of stations, but the multiplicative mechanism can introduce infinite waiting times when CW_{max} is infinite and the number of stations becomes large. Therefore we recommend the use of an additive mechanism that distributes more evenly the channel, and has finite waiting times. Experience has shown that even with limited CW_{max} those two protocols behave better than the IEEE 802.11b one, to the goodput point of view and waiting time one. The additive mechanism keeps remarkably well the optimal saturation throughput of the channel when the number of stations increases.

Many questions remain open. Space is lacking

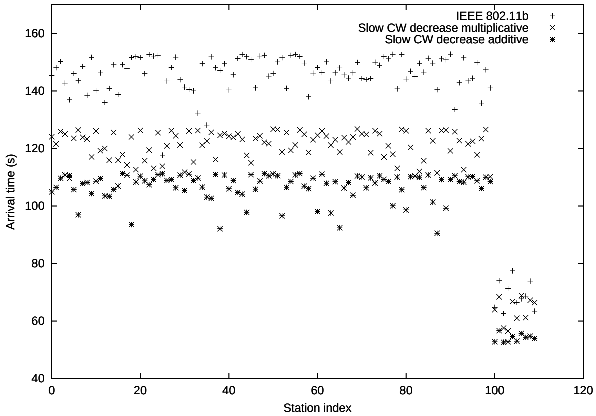

Figure 9: Completion times for the stations with the basic mechanism.

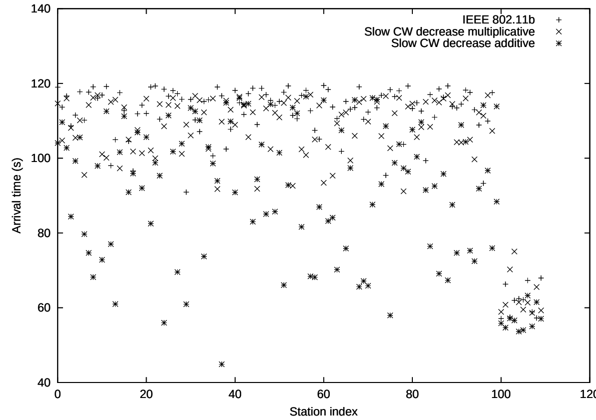

Figure 10: Completion times for the stations with the RTS/CTS mechanism.

here to deal with the fairness question. Other mechanisms in the same family ($g > 1$, $W_i = W_0 + i^2\omega$) seem also promising to many respects. Hidden terminal questions, of course, remain an important issue.

References

[1] Higher-speed physical layer extension in the 2.4 ghz band. IEEE Std 802.11b-1999 Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications.

[2] S. Beuerman and E. Coyle. The delay characteristics of CSMA/CD networks. *IEEE Transactions on Communications*, 36(5):553–563, May 1988.

[3] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(8):535–547, March 2000.

[4] G. Bianchi, L. Fratta, and M. Oliveri. Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs. In *Proc. PIMRC*, pages 392–396, Taipei, Taiwan, October 1996.

[5] A. Branchs, X. Perez-Costa, and D. Qiao. Providing throughput guarantees in IEEE 802.11e wireless LANs. In *Proceedings of ITC'18*, 2003.

[6] 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*, 8(6):783–799, December 2000.

[7] D.-M. Chiu and R. Jain. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. *Computer Networks and ISDN Systems*, 17(1):1–14, June 1989.

[8] S. Garg, M. Kappes, and A.S. Krishnakumar. On the effect of contention-window sizes in IEEE 802.11b networks. Technical Report ALR-2002-024, Avaya Labs, June 2002.

[9] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda. Performance anomality of 802.11b. In *Proceedings of 22nd Conference of the IEEE Communications Society (INFOCOM)*, 2003.

[10] L. Kleinrock and F. Tobagi. Packet switching in radio channels: Part I - carrier sense multiple access modes and their throughput-delay characteristics. *IEEE Transactions on Communications*, 23(12):1400–1416, December 1975.

[11] Q. Ni, I. Aad, C. Barakat, and T. Turlletti. Modeling and analysis of slow CW decrease for IEEE 802.11 WLAN. In *PIMRC*, Beijing, China, September 2003.

[12] S. M. Ross. *Stochastic processes*. J. Wiley & Sons, New York, 1983.

[13] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement. In *Proceedings of INFOCOM'02*, 2002.

[14] Y. Xiao. Saturation performance metrics of the ieee 802.11 mac. In *Proc. of The IEEE Vehicular Technology Conference (IEEE VTC 2003 Fall)*, pages 1453–1457, Orlando, Florida, USA, Oct 2003.