

802.11e EDCA Protocol Parameterization: A Modeling and Optimization Study

Ioannis Koukoutsidis
FORTH-ICS
P.O. Box 1385, 71110 Heraklion, Greece
jkoukou@ics.forth.gr

Vasilios A. Siris
FORTH-ICS/University of Crete
P.O. Box 1385, 71110 Heraklion, Greece
vsiris@ics.forth.gr

Abstract

The IEEE 802.11e MAC protocol specifies an enhanced distributed channel access (EDCA) mechanism with adjustable protocol parameters, providing differentiated access to wireless stations. The EDCA parameters are: CW_{min} and CW_{max} (minimum and maximum contention window), AIFS (arbitration interframe space), and TXOP (transmit opportunity). In this paper, we study the joint setting of these parameters in order to maximize capacity and optimize performance under QoS constraints. We first revisit the analytical modeling in saturation and non-saturation conditions and provide more detailed approximations of the mean channel access delay and throughput. Then, considering two classes of wireless stations with higher and lower QoS demands, and optimizing with respect to average measures and constraints, we provide Pareto optimal pairs for the number of supported stations from the two classes for different parameter set configurations and representative load values. Further, we examine optimal parameter selection for a class of elastic traffic, in the presence of a delay-sensitive class whose parameters are fixed. Our findings show that drastic service differentiation can undermine capacity of a WLAN. They also demonstrate the optimality of jointly setting high values for TXOP and AIFS in order to maximize the throughput of the elastic traffic class while guaranteeing delay of the delay-sensitive class. We also reveal different optimal settings for CW_{min} for different load conditions. We summarize our findings as guidelines for the setting of 802.11e parameters in a scenario with data and real-time service classes.

1. Introduction

The IEEE 802.11 series gives protocol specifications for the interconnection of telecommunications equipment in a WLAN, using a CSMA/CA medium sharing mechanism [4]. In a heterogeneous environment with multiple traffic classes having different service requirements, it is appropriate to provide service differentiation through the channel

access mechanism. With this intent, in the IEEE 802.11e EDCA mechanism finalized in [5] a number of protocol parameters can be set differently for each class of wireless stations.

For a given class i , adjustable EDCA parameters are:¹

- $CW_{min,i}$, $CW_{max,i}$: The minimum and maximum contention windows. Before each transmission attempt and in order to reduce the probability of a collision, a station enters a “backoff stage”, which consists of waiting a random number of time slots, wherein the medium is sensed idle. This number is initially selected uniformly in a minimum “contention window” interval $[0, CW_{min,i} - 1]$. In the event of a collision, a new backoff stage is entered and a larger contention window is selected. The protocol suggests doubling the window until a maximum stage m_i , for which $2^{m_i} \cdot CW_{min,i} = CW_{max,i}$. The number of backoff slots is in general chosen in the interval $[0, 2^{(k \wedge m_i)} \cdot CW_{min,i} - 1]$, where $k = 0 \dots m_i$ is the current backoff stage.
- $AIFS_i$: The arbitration inter-frame space. It is the amount of time a station must sense the channel idle before decrementing its backoff counter, or attempting a transmission. Assigning different values of AIFS to different classes can distribute contention for the channel and prioritize access. For two stations of classes k and j , with class k having a lower priority, the value $(AIFS_k - AIFS_j)$, yielding the AIFS separation in slots, is of practical significance.
- $TXOP_i$: The transmit opportunity. It is used to permit consecutive frame transmissions by a station. It is defined as the maximum interval of time during which a station has the right to initiate a sequence of frame transmissions uninterrupted by others, after it has gained access to the channel through the contention mechanism. It is a basic building block of the

¹These parameters, when used as mathematical variables in this paper, are *italicised*.

enhanced protocol, and every contention win is considered to grant a TXOP. The duration in which the TXOP holder maintains uninterrupted use of the medium includes the last frame’s ACK reception time, therefore it refers to an integral number of frames.

The set $\{CW_{min,i}, CW_{max,i}, AIFS_i, TXOP_i\}$ for each class i is called the *EDCA parameter set*, and is transmitted in the beacon frame of an access point (AP).

The qualitative and quantitative differentiation accomplished through EDCA parameters is fairly known from existing modeling and simulation works [8–10]. The ultimate goal, however, is to find sets of parameters that satisfy given QoS constraints for each class and optimize its performance – or equivalently its capacity, defined as the maximum number of supported stations – under these constraints.

Previous research in this direction tries to accomplish this by evaluating performance or capacity when modifying a subset of the parameters in a given scenario. We discuss some representative works. In [3] the voice capacity of a WLAN is examined, with and without the integration of data traffic, for different settings of the CW_{min} , CW_{max} , and AIFS separation; the authors attest the improvement in performance of voice stations by setting small CW_{min} and CW_{max} values, as expected for unsaturated sources. In [10] the authors argue that, in view of an unsaturated real-time class and a saturated data traffic class, the AIFS separation should be tuned to satisfy delay guarantees of the real-time class, and the CW_{min} of the data class should be adjusted to maximize its throughput. Finally, in [2] the authors examine a traffic scenario with TCP data stations, prioritizing the bottlenecked AP by setting for it a small CW_{min} and increasing the AIFS of data stations.

As a general remark, results so far have been limited to a subset of parameters in specific test cases, and no consistent optimization study has been done. In this paper we attempt such a study, covering *all* EDCA parameters and in representative ranges of loads from low traffic to saturation conditions. We look at a scenario with two service classes, one of which has tighter QoS constraints than the other. Loosely, these refer to delay-sensitive (e.g., voice, streaming video) and elastic (data) traffic. Constraints are in terms of the mean channel access delay and throughput, which are also the measures to be optimized. We provide Pareto optimal pairs for the number of supported stations from the two classes when modifying all parameters. Further, we examine optimal parameter selection for the class of elastic traffic when parameters of the delay-sensitive class are fixed.

From an applications’ viewpoint, our main contributions are the following. We demonstrate the optimality of jointly setting high values for TXOP and AIFS in order to maximize the throughput of the elastic traffic class while guaranteeing delay of the delay-sensitive class. Further, we reveal that CW_{min} should be set differently for different load

conditions, and primarily adopt small values not only when stations are unsaturated, but also when both classes are saturated. It is also shown that having a drastic service differentiation in saturation conditions may seriously undermine capacity of a WLAN when both classes have QoS constraints. In contrast, since constraints are easier satisfied in non-saturation, it is easier to create a tolerable service deterioration in order to improve capacity. To our knowledge, these are the first general conclusions regarding capacity and performance optimization in the presence of both delay-sensitive and elastic traffic, for all EDCA parameters and different loads.

This paper is structured as follows. In Section 2 we describe basic elements of the mathematical modeling of an 802.11e WLAN. Performance measures of interest, namely the average channel access delay and throughput, are calculated in Section 3 for saturation conditions. The approach taken for modeling non-saturation conditions is presented in Section 4. In both Sections 3 and 4 we introduce several refinements compared to previous works. The considered optimization problems are presented and solved in Section 5. Finally, guidelines on the setting of 802.11e parameters for an elastic and real-time class are given in the concluding Section 6.

2. Basic mathematical model

Valuable references for the mathematical modeling of the basic 802.11 protocol are the paper of Bianchi [1] as well as the work of Kumar *et al.* [7]. Appropriate modifications for 802.11e were given in [8,9] for saturation conditions, and presented in the next subsections. Non-saturation conditions will be treated separately in Section 4.

We consider more generally a number of N_i contending stations belonging to service class i ($i = 1, \dots, K$). The stations will be assumed to send traffic to an AP (and not to each other), whose sole purpose is to send back ACKs. Furthermore, we do not consider different traffic classes within the same station. This elementary multiservice model is appropriate for our purposes. Nevertheless, extensions to cover the inclusion of an access point and multiple service classes within one station are straightforward.

2.1. Solving for attempt and collision probabilities

The key element of the mathematical model is the so-called “decoupling approximation” [7]. It purports that the probability of an attempt by a station at each idle-sensed slot, or equivalently the probability of a collision at an attempt is the same throughout the time evolution of the system. As a consequence, the (re)-transmission processes of the different stations are mutually independent (decoupled).

This approximation becomes more accurate as the number of stations increases. Practically, it works well even for a number of stations as low as 2 (see [1]).

Since the collision probability is dependent on the attempt rate and vice-versa, appropriate expressions can construct a system of equations to solve for these values.

Denote the collision and attempt probabilities of a station in class i ($i = 1, \dots, K$) by c_i, p_i respectively. From a stochastic analysis (either a Markovian analysis as in [1], or a renewal theory analysis as in [7]), the attempt rate can be expressed as a function of the collision probability and the backoff parameters $CW_{min,i}, m_i$. Assuming for simplicity that there does not exist a limit on the number of retries to send a packet (see also Remark 2.1), we have

$$p_i = \frac{2(1 - 2c_i)}{(CW_{min,i} - 1)(1 - 2c_i) + CW_{min,i}c_i(1 - (2c_i)^{m_i})} \quad (1)$$

Then, in the frame of the decoupling approximation, considering a population N_i for each class i , we write

$$c_i = 1 - (1 - p_i)^{N_i - 1} \prod_{j \neq i} (1 - p_j)^{N_j} \quad (2)$$

For K classes, we end up with a system of $2K$ nonlinear equations, through which p_i, c_i can be derived numerically.

Remark 2.1 *In the single-class case, the parameter configuration which minimizes the average time between successful transmissions also yields the highest expected number of successes in any bounded interval $(0, t]$, and hence is optimal in any limited or unlimited-retry case. For multiple classes a minor influence can be expected, becoming negligible as the retry-limit increases.*

2.2. Contention zones

Enhancing the idle sensing time of some stations by controlling the AIFS parameter is a basic tool for setting priorities. Consider classes indexed according to “service privilege” order, i.e., $AIFS_1 < AIFS_2 < \dots < AIFS_K$. This creates similarly indexed “zones” of channel activity, where in zone i only classes $j \leq i$ are allowed to contend.

Denote by π_i the stationary probability that the system is in zone i . This can be easily derived by Markov chain analysis (see e.g. [8] for the case of 2 classes). Then the collision probability of a station of class i is

$$c_i = \frac{\pi_i}{\sum_{j=i}^K \pi_j} (1 - (1 - p_1)^{N_1} \dots (1 - p_i)^{N_i - 1}) + \dots + \frac{\pi_K}{\sum_{j=i}^K \pi_j} (1 - (1 - p_1)^{N_1} \dots (1 - p_i)^{N_i - 1} \dots (1 - p_K)^{N_K}), \quad (3)$$

i.e. the sum of collision probabilities for class i in zones $j \geq i$, weighted by the probability of being in each zone when there is an attempt. Along with (1), we have again a system of $2K$ nonlinear equations.

3. Performance measures

The major performance measures we are preoccupied with are the channel access delay and throughput, for each class station. The analysis follows previous works, mainly [3]. It also includes the use of TXOP, which has not been covered in many previous works on 802.11e (including [2, 3, 8, 10]). We only consider basic channel access (i.e., without the use of RTS/CTS [4]), without loss of generality for our results.

The calculation of performance measures is confined to the case of 2 service classes A and B, of which class A will be favored by service differentiation. The following refinements in the analytical model are introduced. We provide a correction in the calculation of the mean channel access delay (and subsequently, throughput) in the case of AIFS differentiation, to include the whole delay a disadvantaged class-B station faces until it is allowed to attempt or perform backoff. Additionally, we calculate the average throughput seen by a station rather than the one seen by the system, quantities which as we show may differ substantially, especially in non-saturation conditions.

3.1. Channel access delay

The channel access delay is defined as the delay a frame experiences from the time it arrives at the head of the transmission queue until it is transmitted successfully, and its transmission is acknowledged. Thus it is (deliberately) set to be a little more than just the “access” time.

We consider a station of class i ($i = A, B$) has a MAC packet size σ_i and transmits at rate R_i . Let also δ be the propagation delay, T_{RxTx} the time for the transceiver to turn around, T_{PLCP} the time to transmit the PLCP preamble and header (adjoined by the physical layer, see [4]) and ack the size of a MAC level acknowledgement. The T_{PLCP} is fixed for each physical layer configuration. Further, ACK packets are always transmitted at a (lower) basic service rate R_b . In addition, a *SIFS* (short inter-frame space) interval is used between the transmission of a frame and the sending of an acknowledgement, to allow the MAC layer to receive the packet and subsequently the transceiver to turn around.

According to the protocol, the duration the medium is busy because of a successful transmission of class i – including the reception of the acknowledgement – is

$$T_i^{succ} = AIFS_i - T_{RxTx} + \delta + T_{PLCP} + \frac{\sigma_i}{R_i} + \delta + T_{PLCP} + SIFS + \frac{ack}{R_b} \quad (4)$$

The time the medium is busy during a collision of class- i stations is

$$T_i^{coll} = AIFS_i - T_{RxTx} + \delta + T_{PLCP} + \frac{\sigma_i}{R_i} . \quad (5)$$

We first derive the mean channel access delay for a station of the priority class A. Following the analysis in [3], for a station of class $i = A$ the channel access delay can be expressed as:

$$D_i^{acc} = \sum_{k=1}^{A_i} \sum_{j=1}^{U^k} S_i^{k,j} + (A_i - 1)T_i^{coll} + T_i^{succ}, \quad (6)$$

where A_i is the number of channel attempts for the given frame, and U^k is a random variable uniformly distributed in $[0, 2^{(k-1) \wedge m_i} CW_{min,i} - 1]$. Under the decoupling approximation, $S_i^{k,j}$ are *i.i.d* random variables representing the duration of each backoff decrement cycle as seen by the station, where in such cycle a successful transmission or collision may follow the idle period. This is referred to as a *generic slot* duration.

The number of transmission attempts follows a geometric distribution with $\Pr\{A_i = k\} = (1 - c_i)c_i^{k-1}$. Packet sizes and transmission rates are deterministic, however the duration of a collision depends on the type of stations implicated. Considering the mean number of backoffs and collisions, we have:

$$\begin{aligned} E[D_i^{acc}] &= E[S_i] E\left[\sum_{k=1}^{A_i} \sum_{j=1}^{U^k} 1_{\{U^k \geq j\}}\right] \\ &+ \frac{c_i}{1 - c_i} E[T_i^{coll}] + T_i^{succ}. \end{aligned} \quad (7)$$

Using that

$$\begin{aligned} E\left[\sum_{j=1}^{U^k} 1_{\{U^k \geq j\}}\right] &= \sum_{j=1}^{2^{k-1} CW_{min,i} - 1} \Pr\{U^k \geq j\} \\ &= \frac{2^{k-1} CW_{min,i} - 1}{2}, \end{aligned}$$

after some calculations we obtain:

$$\begin{aligned} E[D_i^{acc}] &= E[S_i] \left[\frac{CW_{min,i}}{2} \cdot \left(\frac{1 - (2c_i)^{m_i}}{1 - 2c_i} + \frac{(2c_i)^{m_i}}{1 - c_i} \right) \right. \\ &\left. - \frac{1}{2(1 - c_i)} \right] + \frac{c_i}{1 - c_i} E[T_i^{coll}] + T_i^{succ}. \end{aligned} \quad (8)$$

The expected duration of a backoff decrement cycle is

$$E[S_i] = \pi_A \cdot E[S_i^A] + \pi_B \cdot E[S_i^B], \quad (9)$$

where $E[S_i^A]$, $E[S_i^B]$ are the expected generic slot durations for the class $i = A$ station, when it is performing backoff,

if it were in zone A or zone B, respectively. Also, the expected duration of a collision is calculated by considering the $\max(T_A^{coll}, T_B^{coll})$ if it is among a class-A and class-B station, or T_A^{coll} or T_B^{coll} , weighted by their respective probabilities.

Starting from (6), we can approximately calculate the variance of the channel access delay. We finally get for the second moment that

$$\begin{aligned} E[(D_i^{acc})^2] &\approx E[S_i^2] \cdot \Omega_i + \Omega_i(\Omega_i - 1)E^2[S_i] \\ &+ \frac{c_i(1 + c_i)}{(1 - c_i)^2} E[(T_i^{coll})^2] + (T_i^{succ})^2 \\ &+ 2E[T_i^{coll}]E[S_i]\Theta_i + 2E[S_i]\Omega_i T_i^{succ} \\ &+ \frac{2c_i}{1 - c_i} E[T_i^{coll}]T_i^{succ}, \end{aligned} \quad (10)$$

where Ω_i , Θ_i are defined in (12), (13), respectively. Relation (10) is an approximation because we have considered in the calculations that $E[\Omega_i^2] = E^2[\Omega_i]$. Finally, the variance is computed as $\text{Var}[D_i^{acc}] = E[(D_i^{acc})^2] - E^2[D_i^{acc}]$.

We now extend the analysis to derive the mean channel access delay of class-B stations, when there is AIFS differentiation. In this case, after each busy medium end a station of class B will go through a number of slots where it is not entitled to transmit.

To include this period, we consider the AIFS separation $\ell = AIFS_B - AIFS_A$ and the discrete-time Markov chain shown in Fig. 1, with probability $q_A = (1 - p_A)^{N_A}$ and an absorbing state ℓ . The state of the Markov chain represents the number of idle slots that have elapsed since the last busy medium end, as indicated by the carrier sense mechanism of a station.

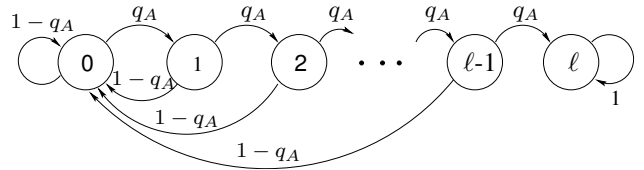


Figure 1. Discrete-time transitions in zone A.

We are interested in the mean number of visits to states $0, 1, \dots, \ell - 1$ prior to absorption in state ℓ . Denote by $\mathbf{P}^n = [p_{ij}^{(n)}]$ ($0 \leq i, j \leq \ell$) the n th power of the transition matrix and its elements. Then, starting from state 0 the mean number of visits to 0 is²

$$m_0 = \sum_{n=0}^{\infty} p_{00}^{(n)}, \quad (11)$$

²For computations, m_0 and $m_{1, \dots, \ell-1}$ are found merely by inverting $(\mathbf{I} - \mathbf{Q})$, where \mathbf{Q} is the restriction of \mathbf{P} to the set of transient states $\{0, 1, \dots, \ell - 1\}$ and \mathbf{I} is an identity matrix with the same dimensions.

$$\Omega_i \triangleq \mathbb{E}\left[\sum_{k=1}^{A_i} \sum_{j=1}^{U^k} 1_{\{U^k \geq j\}}\right] = \frac{CW_{min,i}}{2} \left(\frac{1 - (2c_i)^{m_i}}{1 - 2c_i} + \frac{(2c_i)^{m_i}}{1 - c_i} \right) - \frac{1}{2(1 - c_i)}, \quad (12)$$

$$\Theta_i \triangleq \mathbb{E}\left[\sum_{k=1}^{A_i} (A_i - 1) \sum_{j=1}^{U^k} 1_{\{U^k \geq j\}}\right] = \frac{(CW_{min,i} - 1)c_i}{2(1 - c_i)} + \frac{CW_{min,i}}{2(1 - c_i)} \left[\frac{c_i - [c_i + (m_i - 1)(1 - c_i)](2c_i)^{m_i}}{1 - 2c_i} + \frac{2c_i(1 - c_i)(1 - (2c_i)^{m_i-1})}{(1 - 2c_i)^2} - c_i \right] + 2^{m_i-1} CW_{min,i} \frac{(c_i^{m_i} - c_i^{m_i+1})(m_i - m_i c_i + c_i) + c_i^{m_i+1}(1 - c_i)}{(1 - c_i)^3} - \frac{c_i(1 + c_i)}{2(1 - c_i)^2}. \quad (13)$$

while the mean number of visits to states $1, \dots, \ell - 1$ is

$$m_{1,\dots,\ell-1} = \sum_{n=0}^{\infty} p_{01}^{(n)} + p_{02}^{(n)} + \dots + p_{0\ell-1}^{(n)}. \quad (14)$$

We know that a visit to one of states $1, \dots, \ell - 1$ lasts one idle slot time, while that to state 0 lasts either the time of a successful transmission or of a collision. Therefore the mean channel access delay for a station of class B will become

$$\begin{aligned} \mathbb{E}[D_B^{acc}] &= [\mathbb{E}[S_B^B] + T_{slot} \cdot m_{1,\dots,\ell-1} + T_{busy\ slot}^{zone A} \cdot m_0] \\ &\cdot \mathbb{E}\left[\sum_{k=1}^{A_B} \sum_{j=1}^{U^k} 1_{\{U^k \geq j\}}\right] + \frac{c_B}{1 - c_B} \mathbb{E}[T_B^{coll}] + T_B^{succ}. \end{aligned} \quad (15)$$

In the above expression, T_{slot} is an idle slot time, while $T_{busy\ slot}^{zone A}$ is the mean duration of a busy slot in zone A. The latter writes as

$$\begin{aligned} T_{busy\ slot}^{zone A} &= \frac{N_A p_A (1 - p_A)^{N_A - 1}}{1 - (1 - p_A)^{N_A}} T_A^{succ} \\ &+ \frac{1 - N_A p_A (1 - p_A)^{N_A - 1} - (1 - p_A)^{N_A}}{1 - (1 - p_A)^{N_A}} T_A^{coll}. \end{aligned} \quad (16)$$

In the calculation of $\mathbb{E}[S_B^B]$ above one must set $AIFS_B = AIFS_A$, since the AIFS separation is now accounted for when calculating the mean time to absorption.

Finally, it is noted that in this case it is extremely difficult to compute the delay variance, since second moments of the number of visits to a state prior to absorption are unknown.

3.1.1. Use of TXOP

As mentioned before, wireless stations can carry out multiple frame transmissions, taking advantage of advertised TXOPs. Let a station of class i transmit η_i frames successively (as derived from the $TXOP_i$ limit), where $\eta_i \in \mathbb{N}$. Transmissions are separated by a SIFS period, to allow transceivers to turn around from ACK receptions [5]. A

randomly chosen frame of class i is found with probability $1/\eta_i$ to be at the head of the transmission queue prior to a TXOP grant, and $(\eta_i - 1)/\eta_i$ to be one of the $\eta_i - 1$ following frames that will be transmitted successively. It is clear that the frame at the head of the queue will undergo the standard transmit procedure, and thus suffer a mean access delay obtainable from the previous analysis, which we call $\mathbb{E}[D_i^{acc} | \eta_i = 1]$; on the other hand, the remaining frames benefit by having a much smaller delay, equal to $d'_i = SIFS + \delta + T_{PLCP} + \sigma_i/R_i$. The expected delay of a class- i frame is then

$$\mathbb{E}[D_i^{acc}] = \frac{1}{\eta_i} \mathbb{E}[D_i^{acc} | \eta_i = 1] + \frac{\eta_i - 1}{\eta_i} d'_i. \quad (17)$$

For the calculation of generic slot times, note that the duration of a successful transmission which includes the ACK reception is now increased to $AIFS_i - T_{RxTx} + \eta_i \left(\frac{\sigma_i}{R_i} + 2\delta + 2T_{PLCP} + SIFS + \frac{ack}{R_b} \right) + (\eta_i - 1) \cdot SIFS$. On the other hand, a collision is supposed to occur on the first transmitted frame, and therefore the expression in (5) is unchanged.

Finally, we can also employ (10) to calculate the delay variance in the case where TXOP is used.

3.2. Throughput

The evaluated throughput is the rate of successfully transmitted MAC-level information per unit of time. Denote it by γ_i for a station of class i . The usual way to derive this (e.g., in [1, 3, 9]) is to consider each end of a generic slot as a renewal epoch and calculate the mean amount of successfully transmitted information over the mean generic slot duration,

$$\begin{aligned} \gamma_i &= (\pi_A \sigma_A p_A (1 - p_A)^{N_A - 1} + \pi_B \sigma_B p_B (1 - p_B)^{N_B - 1} \\ &\quad (1 - p_A)^{N_A}) / (\pi_A \mathbb{E}[S_A] + \pi_B \mathbb{E}[S_B]). \end{aligned} \quad (18)$$

However, this can only be characterized as the individual throughput *seen by the system*. To calculate the actual

throughput of a station we have to take the total channel access time as the renewal period, hence

$$\gamma_i = \frac{\sigma_i}{E[D_i^{acc}]} \cdot \quad (19)$$

This further allows to include the corrected calculation of class-B access delay in case of AIFS differentiation.

4. Non-Saturation conditions

In constructing a model for non-saturation conditions, we would like to include the traffic arrival characteristics in its parameters, and simply extend the set of nonlinear equations. The key modeling assumption here is to consider a constant busy station probability at each idle-sensed slot, equal to the load of the station envisaged as a single server queue, as shown in [3].

Consider an arrival rate λ_i for each station in class i , and the queue load ρ_i . The collision probability now is

$$c_i = 1 - (1 - \rho_i p_i)^{N_i - 1} \prod_{j \neq i} (1 - \rho_j p_j)^{N_j} \quad , \quad (20)$$

where the probability of an attempt by class i is conditioned on having a packet to transmit and hence given by the same function of c_i as in the saturation case (1). Along with $\rho_i = \lambda_i \cdot E[D_i^{acc}]$, for K service classes we now have a system of $3K$ nonlinear equations to solve for p_i , c_i , ρ_i .

The mean channel access delay is then calculated following the same approach, which yields a sufficiently good approximation [6]. To calculate the average throughput experienced by a station, we would have to consider a sequence of alternating ON/OFF periods, where the OFF period is geometrically distributed with parameter ρ_i (and may take the value 0). Treating this as a regenerative process, we would have $\gamma_i = \sigma_i / (E[D_i^{acc}] + (1 - \rho_i) / \rho_i \cdot E[S_i])$. This turns out to be extremely inaccurate, yielding throughput results of about 1 order of magnitude greater.

To tackle this, we consider a more involved ON/OFF model where a station may send 1 or more frames successively in the ON period (Fig. 2).

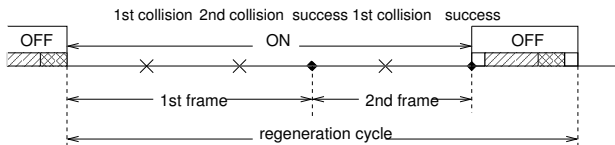


Figure 2. Sample path evolution of the system with ON and OFF periods.

Define r_i^{ON} : the probability of an empty station after a frame transmission and r_i^{OFF} : the probability of an empty

station after a generic slot in the OFF period. We consider these constant and approximate them for Poisson arrivals as

$$r_i^{ON} = e^{-\lambda_i E[D_i^{acc}]} \quad , \quad (21a)$$

$$r_i^{OFF} = e^{-\lambda_i E[S_i]} \quad . \quad (21b)$$

Eq. (21a) represents a low load approximation where consecutive frame transmissions are less frequent; this is because we confine the probability of no arrival in $[0, E[D_i^{acc}])$ (for consecutive transmissions it should be greater).

It is clear that this model describes a regenerative process, as the end of an OFF period is a regeneration epoch. The throughput for class i is calculated as the mean MAC-level information transmitted in a regeneration cycle, over the duration of this cycle:

$$\gamma_i = \frac{\sigma_i / r_i^{ON}}{E[D_{acc,i}] / r_i^{ON} + E[S_i] / (1 - r_i^{OFF})} \quad . \quad (22)$$

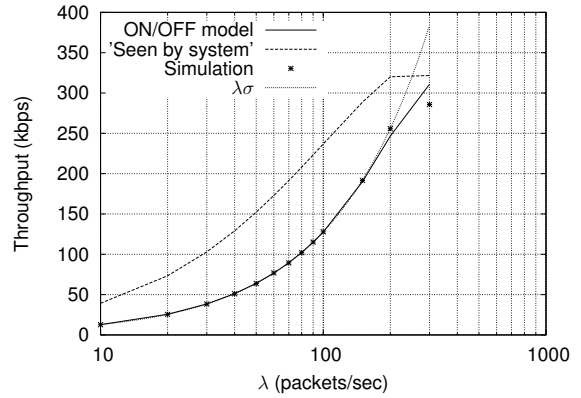


Figure 3. Comparison of different throughput approximations in a 1-class case.

Fig. 3 shows an example of this throughput approximation in a 1-class case with $CW_{min} = 32$, $m = 5$ and $N = 10$ stations, in 802.11a at 6 Mbps, for different Poisson arrival rates until saturation. Results are compared against a discrete-event simulator written in C++. It can also be seen that the corresponding individual throughput “seen by the system” is largely inaccurate for non-saturation conditions (even if the ‘kbps’ throughput scale is a coarse scale when protocol timings are in μs). Also shown on the graph is that the approximation by $\lambda \cdot \sigma$ is extremely accurate for a very large range of loads, which reflects the simple fact that in non-saturation conditions where collisions are fewer, the throughput is almost equal to the information arrival rate.

5. Optimization

We examine two optimization problems in 802.11e. We first assess the capacity of the network for different param-

eter settings, and secondly we attempt to jointly optimize parameters for the two classes A and B, transmitting delay-sensitive and elastic traffic, respectively. In the first problem, we consider delay and throughput constraints for both classes; in the second, a delay bound for class A, while for class B a better-than-best-effort behavior, in terms of the best achievable throughput.

The results are based on the analytic evaluation of mean performance measures in the previous sections. Results in all cases are for 802.11a MAC and PHY layer characteristics. Values of related parameters used in our analytical formulas are summarized in Table 1. The transceiver turnaround time and propagation delay are negligible and are omitted.

Table 1. 802.11a MAC and PHY parameters.

Parameter	Value
T_{slot}	9 μ s
$SIFS$	16 μ s
$\min_i AIFS_i$	34 μ s
T_{PLCP}	20 μ s
ack	14 bytes

We focus on EDCA parameter differentiation and do not introduce other biases. In this sense we set physical transmission rates equal, $R_A = R_B = 6$ Mbps. The basic service rate is also chosen as $R_b = 6$ Mbps. Moreover, the size of MAC packets is set equal to 160 bytes for all stations.³

5.1. Pareto optimal pairs

We seek the maximum number of stations from each class that can be admitted in the system subject to QoS constraints. Since we have contending stations with contradicting performance objectives, we shall derive Pareto optimal pairs⁴. Different parameter settings $\{CW_{min,A}, CW_{min,B}, m_A, m_B, \ell, \eta_A, \eta_B\}$ are examined. Optimal pairs are shown in Fig. 4 for some cases of saturation and non-saturation conditions. In the saturation case, constraints are set as follows: for delay, $E[D_A^{acc}] \leq 5$ ms, $E[D_B^{acc}] \leq 10$ ms, and for throughput, $\gamma_A \geq 300$ kbps, $\gamma_B \geq 200$ kbps. In the non-saturation case we consider more tight constraints, set to be the same for both classes $i = A, B$: $E[D_i^{acc}] \leq 1$ ms, and $\gamma_i \geq 100$ kbps.

The arrangement for this set of results is as follows. We take a “balanced” configuration where EDCA parameters are the same for both classes. These are $CW_{min,A} = CW_{min,B} = 16$, $m_A = m_B = 5$, $\eta_A = \eta_B = 1$, and also

³Bear in mind that the majority of data packets sent over the Internet are also of small size.

⁴A vector (N_A^*, N_B^*) is said to be Pareto optimal iff any other vector (N_A, N_B) in the feasible set has $N_A \leq N_A^*$ or $N_B \leq N_B^*$.

$AIFS_B - AIFS_A = 0$. Subsequently we modify each parameter separately – favoring class A – and derive Pareto pairs. We say a capacity improvement exists in the changed parameter configuration if the Pareto pairs lie “above” those of the balanced case (i.e., in a vector inequality sense). Each modified parameter is shown in the title of each subgraph, and Pareto optimal pairs are depicted by ‘o’ in the unbalanced cases, and by ‘+’ in the balanced ones.

The general goal is to increase capacity of the system by service differentiation in favor of class A, since class-B stations can tolerate lower quality. An overall capacity improvement can be seen in saturation conditions (Fig. 4(a)) in cases where QoS deterioration for class B is more tolerable (cases where $m_B = 10$ and $AIFS_B - AIFS_A = 2$). However, a noteworthy observation is that a more drastic service differentiation can have an adverse effect: when a number of inferior-class stations transmits in the system and requires a certain – even inferior – QoS, capacity of the favored class should largely decrease to accommodate these stations in the network. For instance in the case where $CW_{min,B} = 32$, when no class-B stations exist in the system, the maximum number of allowed class-A stations is 10. When at least 1 class-B station should be able to transmit, the capacity of class A drops to 6. These phenomena are more likely to occur in a saturated network: in non-saturation conditions constraints are easier satisfied and service deterioration usually effects an overall increase in capacity, even with the same constraints for both classes (Fig. 4(b)).

It is worth stressing that the influence of backoff protocol parameters on performance reflects also their influence on capacity. Further results attest that in the protocol, CW_{min} , AIFS and TXOP are more influential parameters than CW_{max} , a parameter which may have no effect at all in non-saturation conditions (see (Fig. 4(b))). The smaller influence of CW_{max} was also witnessed in [10].

5.2. Optimal parameter selection

Parameter design involves searching for performance-optimizing parameters in a space of allowed values. Here we confine ourselves to the finding of optimal parameters for the class of elastic traffic, when the delay-sensitive class has its parameters fixed. This simplifies the automated search and allows for clearer conclusions. Also, in a practical WLAN scenario stations transmitting delay-sensitive traffic are usually fewer in number and unsaturated, and hence little is to be gained further by optimizing their parameters.

An upper delay bound is set for class-A stations. For the data transmitting stations no constraints are imposed, yet we aim at maximizing their throughput. This formulation is consistent with intrinsic QoS demands of delay-sensitive

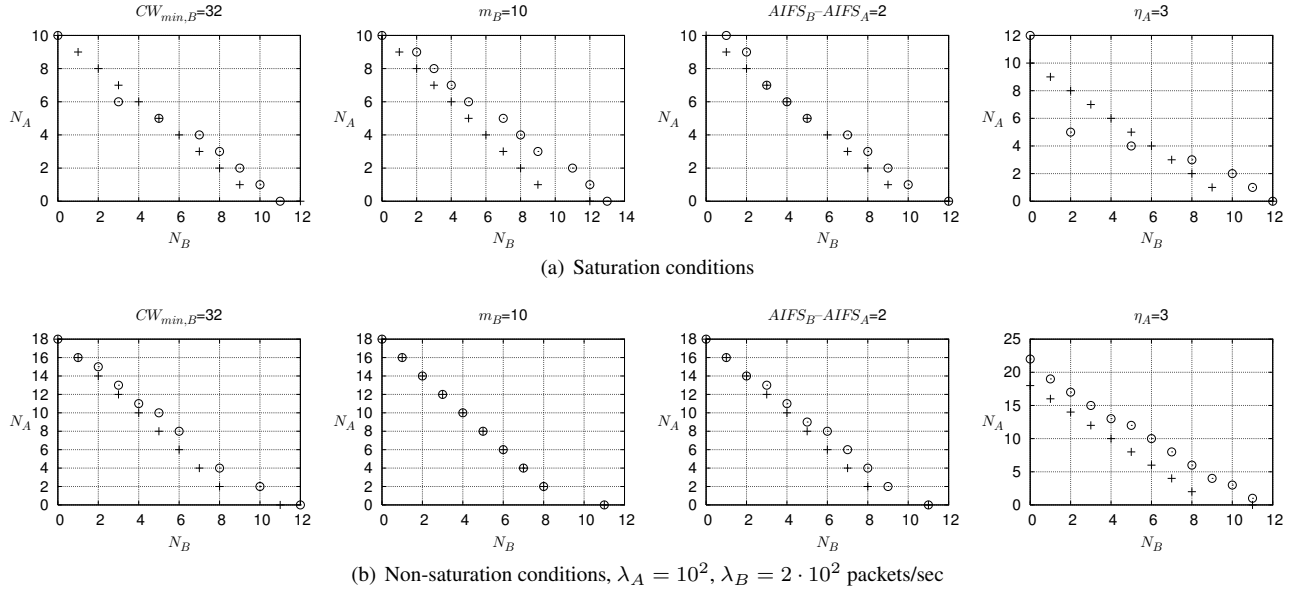


Figure 4. Pareto optimal pairs for various parameter sets.

and elastic traffic. Depending on the number of class-A and class-B stations, we shall find the optimal protocol parameters to set for class B.

We must consider a limited space in our search. In the example we solve, this is the Cartesian product of the value sets that follow, for each parameter:

$$\begin{aligned}
 CW_{min,B} &: \{4, 8, 16, 32, 64, 128, 256\} \\
 m_B &: \{2, 3, 4, 5, 6, 7, 8, 9, 10\} \\
 \ell &: \{0, 1, 2, 3, 4, 5\} \\
 \eta_B &: \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}
 \end{aligned}$$

Class-A parameters are fixed at values $CW_{min,A} = 16$, $m_A = 5$, and $\eta_A = 1$. Results are presented in Table 2, for a case where both classes are saturated, as well as for a case where only class-B stations are saturated, a scenario likely to be encountered in practice. The behavior, with regard to the most influential parameters, is as follows.

- For the case where only class-B stations are saturated, the situation resembles the one where only a single class exists (class-A queue utilization was about 30% in the results of Table 2(b)). Hence, as in the 1-class case, there is an intermediate value of $CW_{min,B}$ which is optimal. However, when both contending classes are saturated, $CW_{min,B}$ should be adjusted to much smaller values, in order to avoid successive channel occupations by the contending class.
- The number of successive transmissions η_B assumes the maximum value in all cases shown in the table.

Hence the exploitation of transmit opportunities is crucial for maximizing the throughput of data traffic.

- Performance degradation of delay-sensitive traffic is at the same time counteracted by increasing the AIFS separation. We see that this is adjusted to higher values for stricter delay constraints (Table 2(a)) or when class-B stations increase in number, in order to prevent the access delay of class A from degenerating inappropriately. Hence in an optimal configuration, depending on load conditions and the tightness of constraints of the delay sensitive class, setting high TXOP value should be supplemented by appropriately increasing AIFS.

The setting of CW_{max} , on the other hand, shows no clear trend, which can be expected since it has a smaller influence on performance. Finally, a useful observation is that the aggregate throughput of elastic traffic increases (with a decreasing rate) as the number of stations increases.

Additional results when also class-A parameters are optimally selected in the same range showed the same trends for class-B parameters, and only small improvements in class-B throughput. Finally, in cases where both classes are unsaturated, it is intuitive that the setting of higher TXOP is less important (since a station has fewer packets to send) and contention windows can be reduced, since there is less congestion in the network.

6. Conclusions and guidelines

Overall the results in this paper have shown that while in the single-class case, given a certain TXOP, an optimal se-

Table 2. Optimal parameter sets

(a) Saturation conditions

Numbers of stations	constraint x ($E[D_A^{acc}] \leq x$ ms)	Optimal parameters for class B				$E[D_A^{acc}]$ (ms)	max γ_B (kbps)	max $N_B \gamma_B$ (Mbps)
		$CW_{min,B}$	m_B	ℓ	η_B			
$N_A = 5, N_B = 1$	5	8	3	2	10	4.760	2774.62	2.775
$N_A = 5, N_B = 10$	5	4	10	3	10	4.933	333.17	3.332
$N_A = 5, N_B = 20$	5	16	5	3	10	4.964	173.43	3.469
$N_A = 5, N_B = 30$	5	8	10	3	10	4.981	117.38	3.521
$N_A = 5, N_B = 1$	3	8	3	4	10	2.915	2074.26	2.074
$N_A = 5, N_B = 10$	3	8	8	5	10	2.867	303.11	3.031
$N_A = 5, N_B = 20$	3	8	8	5	10	2.946	163.25	3.265
$N_A = 5, N_B = 30$	3	8	9	5	10	2.958	112.06	3.362

(b) Class B: saturated, Class A: unsaturated, $\lambda_A = 10^2$ packets/sec

$N_A = 5, N_B = 1$	3	16	2	3	10	2.629	3451.29	3.451
$N_A = 5, N_B = 5$	3	32	10	5	10	2.493	724.57	3.623
$N_A = 5, N_B = 10$	3	32	10	5	10	2.794	371.61	3.716
$N_A = 5, N_B = 20$	3	128	2	4	10	2.871	188.05	3.777

lection of CW_{min} practically suffices to achieve best capacity and performance, in a multiple-class case with different class objectives additional settings are in order.

In the most important scenario with an elastic and real-time class, the following general guidelines can be deduced: Throughput maximization of the elastic class can be achieved by exploiting TXOPs, while delay constraints of the real-time class can be met by increasing the AIFS separation of the two classes. The CW_{min} of the elastic traffic class can be set to its respective 1-class optimal value if it faces light real-time traffic, while for increased real-time traffic it should be adjusted to smaller values. Specific values can be investigated for each problem configuration, based on the analytic modeling presented. Finally, CW_{max} is the least influential parameter and may be fixed at a constant value.

The following critical issues should be considered: First, that increasing the number of frames sent successively via TXOP increases delay variance, since the initial access delay of a frame is comparatively very large. A large transmission variance can result in losses due to retransmission timeouts, even in the presence of a higher-layer protocol with dynamic flow control (e.g., the self-clocking mechanism of TCP). The use of TXOPs is more appropriate for the transmission of larger upper-layer segments, that would anyway be fragmented by the 802.11 MAC and can now be transmitted successively. Secondly, as it was shown in Section 5.1, setting drastic service differentiations should be avoided, since it may seriously undermine the overall capacity of the network.

References

- [1] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J. Select. Areas Commun.*, 18(3):535–547, Mar. 2000.
- [2] P. Clifford, K. Duffy, J. Foy, D. Leith, and D. Malone. Modeling 802.11e for data traffic parameter design. In *Proc. IEEE WiOpt 2006*, Boston, MA, USA, Apr. 2006.
- [3] N. Hegde, A. Proutière, and J. Roberts. Evaluating the voice capacity of 802.11 WLAN under distributed control. In *Proc. IEEE LANMAN 2005*, Chania, Greece, Sept. 2005.
- [4] IEEE Computer Society. *ANSI/IEEE Std 802.11*, 1999.
- [5] IEEE Computer Society. *IEEE Std 802.11e-2005*, 2005.
- [6] I. Koukoutsidis and V. Siris. Modeling approximations for an IEEE 802.11 WLAN under Poisson MAC-level arrivals. In *Proc. IFIP Networking '07*, Atlanta, GA, USA, May 2007.
- [7] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs. In *Proc. IEEE Infocom '05*, pages 1550–1561, Miami, FL, USA, Mar. 2005.
- [8] V. Ramaiyan, A. Kumar, and E. Altman. Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness, multi-stability, and throughput differentiation. In *Proc. ACM Sigmetrics '05*, pages 109–120, Banff, Canada, June 2005.
- [9] J. Robinson and T. Randhawa. Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function. *IEEE J. Select. Areas Commun.*, 22(5):917–928, June 2004.
- [10] I. Tinnirello, G. Bianchi, and L. Scalia. Performance evaluation of differentiated access mechanisms effectiveness in 802.11 networks. In *Proc. IEEE Globecom '04*, volume 5, pages 3007–3011, Dallas, TX, USA, Nov. 2004.