A Novel Effective Capacity-Based Framework for Providing Statistical QoS Guarantees in IEEE 802.11 WLANs☆

Emmanouil Kafetzakis[a,b], Kimon Kontovasilis[a,*], Ioannis Stavrakakis[b]

[a] *National Center for Scientific Research "Demokritos", Inst. Informatics & Telecommunications, GR-15310, Ag. Paraskevi, Greece.*
[b] *National & Kapodistrian University of Athens, Dept. Informatics & Telecommunications, Panepistimiopolis, Illisia, Athens 15784, Greece.*

Abstract

This article proposes a performance model of the IEEE 802.11 MAC layer that employs the notion of Effective Capacity. In particular, the paper establishes that an IEEE 802.11 mobile station can be regarded as a Semi-Markovian bursty server of the On/Off type, with known distributions for the On and Off periods, and subsequently applies known results for Semi-Markovian models to derive the Effective Capacity function of this On/Off server. The general Effective Bandwidth/Capacity theory can then be used for computing buffer overflow probabilities and for employing simple traffic control policies to enforce related QoS guarantees. The policies guarantee a soft bound on the buffer overflow probability and are suitable for real-time traffic control over WLANs. The Effective Capacity model of IEEE 802.11 stations is originally developed by assuming that the other competing stations are saturated. This is a conservative assumption that becomes very accurate in a highly loaded network. Subsequently, the model is adapted to encompass lightly loaded networks as well. In the adapted model, each mobile station directly measures a few model parameters, instead of calculating them on the basis of the saturation assumption, and uses these measurements in the computation of its Effective Capacity function. The theoretical results are checked against simulations, validating the appropriateness of the model.

Keywords: Effective Bandwidth, Effective Capacity, IEEE 802.11, QoS, Admission control, tail-probabilities.

1. Introduction

Wireless networking has gained widespread acceptance, because it allows users to connect and exchange information flexibly with no need for cabling. Since IEEE 802.11 is the prevalent standard for Wireless LANs (WLANs), it has been studied extensively under various traffic loads and wireless channel conditions [1–11]. As a general remark, the performance evaluation of the IEEE 802.11 protocol is a somewhat involved problem, due to the complexity of the backoff mechanism in the Medium Access Control (MAC) layer and the associated interdependencies between competing mobile stations.

Reference [1] considers a p-persistent model for modeling and optimizing the performance of IEEE 802.11; in that model, backoff intervals are sampled from a geometric distribution. Along a different approach, and in order to estimate the throughput under the assumption that every station has always a packet to send (saturation condition), [2] models the IEEE 802.11 Distributed Coordination Function (DCF) using a two-dimensional Markov chain to represent the backoff dynamics of each station. The work in [2] has served as a starting point for many subsequent studies; for example, [3] explores service differentiation through the use of different system parameter values. More recently, [2] was generalized and supplemented in [4], in a way that permits the usage of arbitrary backoff window distributions.

☆The first and second authors' work was funded in part by the European Union through the UNITE project, under contract IST-FP6-STREP-026906. The first and third authors' work was funded in part by the European Union through the NoE CONTENT, under contract IST-FP6-384239.

*Corresponding author

Email addresses: mkafetz@iit.demokritos.gr (Emmanouil Kafetzakis), kkont@iit.demokritos.gr (Kimon Kontovasilis), ioannis@di.uoa.gr (Ioannis Stavrakakis)

The mean value of the access delay to the saturated medium is calculated in [4–6], while [7–10] consider higher moments as well. The study in [7] assumes that the access delay under saturation conditions is a gaussian random variable and computes the probability that the backoff delay is below a given threshold. Reference [8] derives the access delay as a function of the number of transmissions that the head-of-the-line packet sees during the station's backoff interval and uses this result for the computation of the access delay probability distribution.

Papers [9, 10] target the computation of queue size-related metrics. These works characterize the access delay using z-transform techniques and then employ a queueing system whose service time features the same probability generating function as the said access delay. The model in [9] and its refinement in [11] employ an infinite G/G/1 queueing system, targeting the calculation of the mean queue size, while [10] relies on a simpler M/G/1/K queueing model towards approximating the queue length probability distribution function. The results of [10] are applicable only to systems featuring Poisson traffic patterns. Furthermore, the approximation of the queue-length distribution requires extensive numerical calculations for obtaining the steady state probabilities of the finite buffer system. Computational complexity issues and numerical precision considerations become more pronounced when large buffers (and small corresponding overflow probabilities) are considered.

None of the results just reviewed are directly applicable to the computation and/or enforcement of QoS related to low probability percentiles. Such tail-related QoS guarantees are becoming increasingly important, since they serve as service requirement descriptors for applications that cannot be accommodated by mean value analysis (see, e.g., some of the QoS specifications in [12]). Generally, when one is interested in stochastic tail-related QoS, asymptotic approaches (most times based on Large Deviations theory) are appropriate. In particular, Effective Bandwidth/Capacity theory is very appealing, because it provides a framework for obtaining asymptotically tight approximations of tail-probabilities and for formulating associated traffic control schemes in a unified way, applicable to arbitrary traffic patterns (provided these patterns possess a well-defined Eff. Bandwidth function). In contrast, conventional queueing theory approaches would require a separate model and perhaps a different methodological approach for every different type of traffic that may be encountered. Although the Eff. Bandwidth/Capacity approach has the shortcoming of being an asymptotic theory, valid to the limit as the system's buffer tends to infinity and the corresponding overflow probability tends to zero, results from many domains of application indicate that this approach is sufficiently accurate for engineering purposes.

With respect to the asymptotic context just outlined, the only relevant works, to the best of the authors' knowledge, are [13] and [14]. Reference [13] tries to guarantee a certain station overflow probability through the application of the Eff. Bandwidth/Capacity theory. The relevant results provide only a rough approximation as they neglect the random variations of the service capacity in IEEE 802.11, simplistically assuming that the said capacity is constant and equal to the difference between saturated and unsaturated throughput (i.e., residual bandwidth). Publication [14] tries to improve on that, by capturing the burstiness of the IEEE 802.11 server. On the basis of the assumption that the IEEE 802.11 DCF function exhibits memoryless behaviour when all competing stations have backlogged packets and the additional strong assumption that all stations in the WLAN feature a homogeneous traffic load, whose profile is restricted to the exponential On/Off type, [14] models the service capacity of each IEEE 802.11 station in the WLAN as a Markov-Modulated Poisson Process (MMPP). The establishment of the MMPP model involves a series of further approximations, towards representing the MAC dynamics in simplified terms. In particular, the model developed in [14] is suitable only for moderate traffic loads, because otherwise (i.e., in either light or heavy loads) several expressions relevant to key model parameters become inaccurate. Although [14] attempts to extend the model to encompass heterogeneous traffic conditions (still restricted only to the exponential On/Off type), this is done in a way that preserves a sense of 'equivalence' in terms of some input traffic parameters, but neglects the impact of the different traffic loads on the MAC dynamics. It is noted that traffic control policies based on [14] inherently assume that each station has a complete knowledge of the traffic load of all competing stations.

Our focus in this paper is different, in the sense that we propose a modeling methodology that allows the accurate calculation of the IEEE 802.11 Effective Capacity as *independently* seen by each competing station, without the requirement for knowledge of traffic details in the WLAN. In fact, the Eff. Capacity approach proposed in this paper is applicable under arbitrary station traffic patterns (provided these patterns possess a well-defined Eff. Bandwidth function). Moreover, admission control tests are performed at each station on the basis of locally available information only. Our modeling approach accurately tracks all important MAC characteristics, in line with state of the art approaches previously employed in throughput computation.

The model is firstly developed on the basis of the assumption that, apart from the observed station, all other competing stations are saturated. This is a conservative assumption which leads to results that are very accurate in highly loaded networks. Subsequently, the saturation assumption is relaxed and the proposed approach is shown to be still applicable, provided that a few simple model parameters are distributively measured by the stations (instead of being calculated on the basis of the saturation assumption). All required parameters can be reliably measured in a short time, since the corresponding events occur frequently under all traffic loads. This adaptation of the model to all network loads still avoids the requirement for any knowledge of input traffic details.

The rest of the paper is organized as follows: Section 2 briefly reviews the background on the Eff. Bandwidth and Eff. Capacity theory, while Section 3 discusses special properties of the Eff. Capacity function of Semi-Markovian server models. The Semi-Markovian setting is directly relevant to the model adopted for the IEEE 802.11 WLAN. This model is developed in Section 4, which establishes that the IEEE 802.11 DCF function can be represented as a Semi-Markovian On/Off data server and derives the Eff. Capacity of the IEEE 802.11 protocol on the basis of this representation. The adaptation of the model to all network load settings is discussed in Section 5, where the assumption of saturated conditions is waived. Subsequently, Section 6 validates the model by comparing analytical and simulation results for the Eff. Capacity function and the overflow probabilities, in both highly and lightly loaded environments. Finally, the article is concluded in Section 7.

2. Background on Effective Bandwidth and Effective Capacity Theory

The Eff. Bandwidth/Capacity theory provides for an asymptotically tight linkage between source characteristics, system resources (i.e., server capacity and buffer size) and QoS. A large number of works have contributed to the development of the theory (see, e.g., [15] for a survey), which was originally applied in the context of wired Asynchronous Transfer Mode (ATM) networks. The Eff. Bandwidth theory encapsulates the traffic details of time-varying bursty sources in a single function, namely the *Effective Bandwidth function*, that can be used to express the minimal server capacity required to satisfy a given overflow probability-related QoS constraint. The theory was originally developed for queueing systems with constant server capacity.

When the server's capacity varies with time independently from the input, the theory can be generalized by defining an *Effective Capacity function* to capture the server's burstiness. This function can be used to estimate the maximal arrival rate that can be processed from the specific server while satisfying a given overflow probability-related constraint. Although this generalization has been studied for some years (see, e.g., [16–18]), it did not attract much attention until recently [19, 20] when the importance of wireless communication systems grew considerably. This is because most such systems feature a variable service rate and thus the notion of Eff. Capacity appears to be ideal for their modeling. In most past works the application of Eff. Capacity to wireless systems has focused on the modeling of the rate fluctuations at the physical layer [19, 20]. Instead, this paper employs the Eff. Capacity theory for MAC layer modeling.

For a quick review of the Eff. Bandwidth/Capacity theory, consider first a queue of constant service rate c, fed by traffic that produces an amount of data $V(t)$ within a time window $(0, t]$. According to the Eff. Bandwidth theory, provided that the process $V(t)$, $t \geq 0$, has stationary and ergodic increments and satisfies some additional mild technical conditions[1] (see, e.g., [16, 18, 21]), the probability that the stationary queue size Q exceeds a certain threshold x has an asymptotically exponential upper bound as $x \to \infty$, if a suitable condition holds. Specifically,

$$a_B(\theta) < c \Rightarrow \lim_{x\to\infty} \frac{-\log \Pr\{Q > x\}}{x} \geq \theta, \tag{1}$$

where

$$a_B(s) \triangleq \frac{1}{s} \lim_{t\to\infty} \frac{1}{t} \log \mathrm{E}\left[e^{sV(t)}\right], \quad s \in \mathbb{R}, \tag{2}$$

is the Eff. Bandwidth function of the said traffic[2]. In words, (1) states that, if the value of the Eff. Bandwidth function for some argument θ is less than the server's capacity, then the queue-length distribution will have an asymptotically

[1] The technical conditions are essentially those required for the applicability of the Gärtner-Ellis' theorem and are always satisfied for the class of models employed in this paper.

[2] The requirement for a finite limiting log-moment generator at the right hand-side of (2) implies that the theory is applicable when the traffic process is free from long-range dependence.

exponential decay at a rate θ or faster. A partial converse to (1) also exists: if $a_B(\theta) > c$, then the tail of the queue-length distribution cannot be exponentially bounded with a decay rate θ.

Due to the convexity of the asymptotic log-moment generator in (2), the Eff. Bandwidth function $a_B(\cdot)$ is increasing, with a range between the mean and peak rate of the input traffic (for nonnegative arguments $s \geq 0$). Obviously, due to (1) and its converse, the tightest exponential decay rate that can be achieved in (1) for a given server capacity c is equal to $\theta^\star = \sup\{\theta \mid a_B(\theta) < c\}$. When $\theta^\star$ is finite (as in the case when $a_B(\cdot)$ is *strictly* increasing and c is between the mean and peak traffic rates, whereas $\theta^\star = a_B^{-1}(c)$), then there is an asymptotic 'match', viz.,

$$\lim_{x\to\infty} \frac{-\log \Pr\{Q > x\}}{x} = \theta^\star. \tag{3}$$

Larger values of $\theta^\star$ indicate faster decay rates (satisfying a stricter QoS constraint), while smaller values of $\theta^\star$ imply slower decay rates (satisfying a looser QoS constraint). For the strictly increasing case, by combining (1), its converse, and (3) one arrives at the equivalence

$$a_B(\theta) \leq c \Leftrightarrow \lim_{x\to\infty} \frac{-\log \Pr\{Q > x\}}{x} \geq \theta. \tag{4}$$

It is mentioned that the Eff. Bandwidth function of a traffic stream does not depend on other independent streams that might be multiplexed with and it is additive. Therefore, the Eff. Bandwidth function of the superposition of a number of independent traffic flows is simply the sum of the Eff. Bandwidths of the constituent flows. Because of this important property, simple traffic control schemes arise naturally, reminiscent of classical circuit-switching.

Now consider a queue with time-varying service capacity, where the capacity fluctuations are independent from the input. Let the traffic be as previously and denote by $C(t)$ the amount of data that the server can process within the time window $(0, t]$. Assuming the same technical conditions for the input and output processes as before, the probability that the stationary queue size Q exceeds a certain threshold x satisfies a relation similar to (1), namely

$$a_B(\theta) < a_C(-\theta) \Rightarrow \lim_{x\to\infty} \frac{-\log \Pr\{Q > x\}}{x} \geq \theta, \tag{5}$$

where $a_B(\cdot)$ is the Eff. Bandwidth function defined in (2) and

$$a_C(s) \triangleq \frac{1}{s} \lim_{t\to\infty} \frac{1}{t} \log \mathrm{E}\left[e^{sC(t)}\right], \quad s \in \mathbb{R}, \tag{6}$$

defines the Eff. Capacity function of the server [17, 18]. Comparison of (5) with (1) reveals that the Eff. Capacity value $a_C(-\theta)$ now acts in place of the previously constant server capacity. Again, a partial converse exists: if $a_B(\theta) > a_C(-\theta)$, the tail of the stationary queue-length distribution cannot be exponentially bounded with a decay rate θ. The reason for which $a_C(\cdot)$ is used with non-positive arguments in (6), in contrast with $a_B(\cdot)$ that employs a non-negative argument, is that the amount of data $C(t)$ that can be served in $(0, t]$ may be regarded as a 'virtual' negative input traffic $-C(t)$, independent from $V(t)$, the superposition of the two being applied to a fictitious work-conserving queue with a constant server rate of value $c = 0$. In terms of this equivalence, and due to the additivity of the Eff. Bandwidth functions, (5) is essentially not different than (1).

The Eff. Capacity function $a_C(\cdot)$ is increasing (because $f_t(s) \triangleq \log \mathrm{E}\left[e^{sC(t)}\right]$ is convex with $f_t(0) = 0$, and these properties pass to the limit), with a range between the minimal and the mean service rate (for nonpositive arguments $s \leq 0$). Similarly to the constant capacity case, the tightest decay rate that can be achieved is equal to $\theta^\star = \sup\{\theta \mid a_B(\theta) < a_C(-\theta)\}$. If at least one of $a_B(\cdot)$, $a_C(\cdot)$ is *strictly* increasing, $a_B(0) < a_C(0)$ (i.e., the system is stable, featuring a mean traffic rate less than the mean server rate) and $\lim_{s\to+\infty} a_B(s) > \lim_{s\to-\infty} a_C(s)$ (i.e., peak input rate greater than the minimal service rate) then instead of an upper bound, we have an asymptotic 'match' [18, 21], i.e., there is a unique positive $\theta^\star$ satisfying

$$a_B(\theta^\star) = a_C(-\theta^\star), \tag{7}$$

and (3) holds.

Given that the above-mentioned *strict* monotonicity holds for at least one of the Eff. Bandwidth and Eff. Capacity functions, we can collectively express the results in (5), its converse, (7) and (3) as

$$a_B(\theta) \leq a_C(-\theta) \Leftrightarrow \lim_{x\to\infty} \frac{-\log \Pr\{Q > x\}}{x} \geq \theta, \tag{8}$$

which generalizes (4). Equivalence (8) always holds in the context of this paper because, as we shall see, the Eff. Capacity of IEEE 802.11 is always a strictly increasing function. Note that sometimes the Eff. Capacity is regarded in terms of the QoS parameter $\theta > 0$, rather than the function's original argument $-\theta < 0$. With respect to the QoS parameter, the Eff. Capacity is a decreasing function.

We close this section with a comment on the provision of stochastic QoS guarantees. Consider a loss-related QoS requirement dictating that the queue content should not exceed some given level x (this event being taken as a proxy to overflows in a finite buffer of size x) with probability greater than $e^{-\epsilon}$. In light of the requirement $\Pr\{Q > x\} \leq e^{-\epsilon}$, the equivalence (8) implies that (asymptotically, for large x) one must have $\theta \geq \epsilon/x$, so the condition $a_B(\epsilon/x) \leq a_C(-\epsilon/x)$ must be satisfied. This condition is directly suitable for use as a traffic admission control test.

3. The Effective Capacity of Semi-Markovian Servers

This section discusses special properties of the Eff. Capacity function of Semi-Markovian server models, by exploiting results of [22] on the closely related problem of Eff. Bandwidths for Semi-Markovian (input) traffic processes. The models under consideration comprise a finite number K of states, each corresponding to a service rate r_j, $j = 1, \dots, K$. Not all rates are equal, otherwise the model would describe a server with constant capacity. Transitions between states occur according to a discrete, irreducible Markov chain, characterized by the transition probability matrix $\mathbf{P}$ and an invariant probability vector $\boldsymbol{\pi}$, satisfying $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\sum_j \pi_j = 1$. The sojourn time for a particular visit to some state is independent of the states visited in the past or to be visited in the future and the corresponding sojourn times. Finally, different visits to a state correspond to independent and identically distributed random variables.

In view of the previous comment, a generic random variable T_j may be used for the sojourn time in state j, for any visit to this state. Use $\gamma_j(\omega) \triangleq \mathrm{E}\left[e^{\omega T_j}\right]$ to denote the moment generator of T_j, with an effective domain $\Omega_j \triangleq \{\omega \in \mathbb{R} \mid \gamma_j(\omega) < \infty\}$, and let $\omega_j^* \triangleq \sup \Omega_j$. Arbitrary distributions are allowed for the variables T_j, the only restrictions being $\omega_j^* > 0$ (i.e., the distribution of T_j is not heavy tailed), and $\omega_j^* \notin \Omega_j$ (open effective domains). Under these restrictions, which will hold throughout this paper, all $\gamma_j(\omega)$ are analytic functions in their entire effective domain $\Omega_j = (-\infty, \omega_j^*)$, $j = 1, \dots, K$. From standard Markov Renewal theory, the mean service rate for this class of models is

$$\bar{r} = \frac{\sum_{j=1}^{K} \pi_j \mathrm{E}\left[T_j\right] r_j}{\sum_{j=1}^{K} \pi_j \mathrm{E}\left[T_j\right]}. \tag{9}$$

Also let $r_{\min} = \min_j r_j$.

Introduce two variables, s and u, consider the values $\gamma_j(r_j s - u)$ for all the combinations of s and u that satisfy $r_j s - u \in \Omega_j$, and let $\mathbf{\Gamma}(s,u) \triangleq \mathrm{diag}\{\gamma_j(r_j s - u)\}$. For each permissible pair of values s and u, the matrix $\mathbf{A}(s,u) \triangleq \mathbf{\Gamma}(s,u)\mathbf{P}$ is a nonnegative irreducible matrix with spectral radius $\phi(s,u) \triangleq \rho(\mathbf{A}(s,u))$. Given the notation just introduced, Theorem 3.1 in [22] assures that for every $s \leq 0$ there exists a unique $u(s)$, such that $\phi(s, u(s)) = 1$. This value satisfies $\bar{r}s \leq u(s) \leq r_{\min}s$.[3] Moreover, the function $u(s)$ is analytic and strictly increasing, featuring $u(0) = 0$ (so $u(s) < 0$, for all $s < 0$) and $u'(0) = \bar{r}$.

Furthermore, Theorem 3.2 in [22] guarantees that the Eff. Capacity function of the Semi-Markovian server model exists and is equal to

$$a_C(s) = u(s)/s, \quad \forall s < 0. \tag{10}$$

By inheriting properties from $u(\cdot)$, it is seen that $a_C(\cdot)$ is an analytic function featuring $r_{\min} \leq a_C(s) \leq \bar{r}$, for all $s \leq 0$. Note that the upper bound is tight since, by L' Hôspital's rule, $\lim_{s \to 0} a_C(s) = u'(0) = \bar{r}$. The Eff. Capacity of non-degenerate Semi-Markovian models (i.e., models that feature at least one state with rate different from the model's mean rate and corresponding sojourn time that is not almost surely constant) is a *strictly* increasing function (see Theorem 3.3 in [22]). The IEEE 802.11 model of Section 4 is non-degenerate.

In light of the results just reviewed, to obtain the value of the Eff. Capacity function for a given argument $s = \theta$, one has to solve for the unique u satisfying $\phi(\theta, u) = 1$. This requires the employment of a numerical zero-finding

[3] Actually this is an immediate variant of Theorem 3.1 in [22], which dealt with Eff. Bandwidths of Semi-Markovian traffic processes. For the latter, $s \geq 0$ and $\bar{r}s \leq u(s) \leq \max_j r_j$.

method. However, when one is interested in admission control, the test in the left hand-side of (8) suffices. We now show that this test is equivalent to simply evaluating $\phi(s,u)$ for suitable s and u, avoiding the need for a zero-finding method.

Proposition 1. *For Semi-Markovian server models, the left hand-side condition of (8) is equivalent to*

$$\phi(-\theta, -\theta a_B(\theta)) \leq 1, \tag{11}$$

for any $\theta > 0$.

Proof. Assume that the left hand-side of (8) holds, viz., $a_B(\theta) \leq a_C(-\theta)$. Then, (10) and the fact that $\theta > 0$ yield

$$u(-\theta) \leq -\theta a_B(\theta), \tag{12}$$

which, combined with the fact that all moment generators $\gamma_j(\omega)$ are increasing functions of their argument, implies

$$\mathbf{\Gamma}(-\theta, -\theta a_B(\theta)) \leq \mathbf{\Gamma}(-\theta, u(-\theta))$$

(matrix inequalities are to be interpreted element-wise). This, in turn, leads to

$$\mathbf{A}(-\theta, -\theta a_B(\theta)) \leq \mathbf{A}(-\theta, u(-\theta)),$$

because $\mathbf{P}$ is nonnegative. From this last relation and the property that the spectral radius of nonnegative matrices is an increasing function of any of their elements (i.e., if $\mathbf{0} \leq \mathbf{A} \leq \mathbf{B}$, then $\rho(\mathbf{A}) \leq \rho(\mathbf{B})$) [23, p. 27], we have

$$\phi(-\theta, -\theta a_B(\theta)) \leq \phi(-\theta, u(-\theta)) = 1, \tag{13}$$

where the final equality follows from the definition of $u(\cdot)$ and proves that the left condition in (8) implies (11).

In the reverse direction, assume (11), which is equivalent to (13). By the arguments previously employed, $\phi(s,u)$ is decreasing in u, and thus (13) implies (12), which is equivalent to the left hand-side condition in (8), because $\theta > 0$. □

For the important case of On/Off servers, the implicit function problem (derived from the requirement $\phi(s,u(s)) = 1$) simplifies further. Let States 1 and 2 correspond to the On and Off periods, respectively, and denote the peak rate of the server by $\hat{r}$. Then, $\mathbf{P} = \binom{0\,1}{1\,0}$ and $r_1 = \hat{r}$, $r_2 = r_{\min} = 0$. Therefore,

$$\mathbf{A}(s,u) = \begin{pmatrix} 0 & \gamma_{\mathrm{on}}(\hat{r}s - u) \\ \gamma_{\mathrm{off}}(-u) & 0 \end{pmatrix}$$

with $\phi(s,u) = \sqrt{\gamma_{\mathrm{on}}(\hat{r}s - u)\gamma_{\mathrm{off}}(-u)}$, where $\gamma_{\mathrm{on}}(\cdot)$ and $\gamma_{\mathrm{off}}(\cdot)$ stand for the moment generators corresponding to the distributions of the On and Off sojourn times, respectively. Since $\phi(s,u(s)) = 1$, a logarithmic transformation suggests that $u(s)$ is the unique negative solution of

$$\log \gamma_{\mathrm{on}}(\hat{r}s - u) + \log \gamma_{\mathrm{off}}(-u) = 0, \tag{14}$$

and $a_C(s)$ is again obtained through (10). Therefore, the relation

$$\log \gamma_{\mathrm{on}}(-\hat{r}\theta + \theta a_B(\theta)) + \log \gamma_{\mathrm{off}}(\theta a_B(\theta)) \leq 0 \tag{15}$$

is equivalent to (11) for the class of On/Off servers.

4. The Effective Capacity of IEEE 802.11 Stations

In 1997, IEEE adopted the IEEE Std. 802.11-1997 [24], which specified the Physical (PHY) and Medium Access Control (MAC) layers. Thereafter, further related IEEE standards were released [25–27], describing different PHY layers. In contrast, the IEEE 802.11 MAC layer remained unchanged until the release of the IEEE 802.11e protocol [28]. The IEEE 802.11 MAC layer has two modes of operation: the Distributed Coordination Function (DCF) for networks without infrastructure (ad-hoc) and the Point Coordination Function (PCF) for access point coordination. In this paper we consider the DCF mode of operation, since the PCF mode is optional and it is not implemented in most commercial products. The DCF mode for sharing access to the wireless medium is based on a Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol with binary exponential backoff. Two access mechanisms exist in the DCF: the Basic Access (two-way handshaking) and the RTS/CTS (Request to Send/Clear to Send) access (four-way handshaking). In the rest of the paper it is assumed that the reader is familiar with the IEEE 802.11 terminology.

The remainder of this section is divided into four parts: Subsection 4.1 deals (by drawing on prior work) with certain key probabilities whose computation is needed as input to the subsequent stages of analysis. Subsection 4.2 contains the main contribution of this paper, namely the development of a simple model, from which the Eff. Capacity of the IEEE 802.11 DCF is derived. Subsection 4.3 shows how the Eff. Capacity derivations of this paper can be employed in a traffic admission control scheme that helps provide stochastic QoS guarantees. The analysis in Subsection 4.2 assumes the RTS/CTS access mode, but the particular access mode in use affects the model only marginally, through the values of a few timing constants. Subsection 4.4 discusses the few modifications that are required for using the IEEE 802.11 Eff. Capacity model with WLANs employing the Basic Access mode.

4.1. Modeling Foundations

A simple but accurate analytic model for the computation of the saturation throughput achieved by the IEEE 802.11 protocol was presented in [2] and its enhancement [4]. In preparing for the analysis leading to the Eff. Capacity model in Subsection 4.2, we now briefly review key concepts and results from these works and introduce relevant notation.

The work in [2] assumes saturation conditions and a given (constant) number of stations n in contention. The probability of a collision experienced by a packet being transmitted on the channel, named conditional collision probability p, is assumed to be constant and independent from the number of retransmissions already suffered. The wireless channel is considered to be error-free and without hidden stations.

In order to compute the station's saturation throughput, [2] studies the behaviour of a station through a discrete-time Markov chain model with states of the form (i,j), signifying that the station is at the i^{th} backoff stage and its backoff counter value is j. Transitions of the Markov chains corresponding to different stations are assumed synchronized, but otherwise different Markov chains evolve independently. The discrete-time Markov chain for a station is a process embedded at (continuous-time) instants when the backoff counters of the listening stations decrease their values. The Markov chain model can be solved for the probability that a station transmits in a random model-slot, referred to as transmission probability τ. As pointed out in [2], the probability τ is independent from the wireless access mechanism (usage or not of the RTS/CTS handshaking).

Reference [4] proposed a more general model, which permits the usage of arbitrary backoff counter distributions, generalizing and supplementing [2]. By employing the more general framework of [4] and allowing for an infinite number of retries, one can construct a Markov chain to model the backoff dynamics of a given station in a way analogous to that proposed in [2]. The state transitions of the Markov chain just mentioned are depicted in Fig. 1, where for the moment the dashed box to the upper left of the figure is to be ignored (i.e., direct transitions from states $(i,0)$, $i = 0,\dots,m$ to states $(0,j)$, $j = 0,\dots,W_0 - 2$ are to be assumed).

A station transmits when its backoff counter value reaches zero. At each transmission attempt, and regardless of the number of retransmissions already suffered, each packet is assumed to collide with probability p. If collision occurs when the station is at the $(i-1)^{\text{th}}$ backoff stage, the collided station moves to the i^{th} backoff stage and draws a new backoff counter value $l \geq 0$ with probability $p_l^{(i)}$. The quantity m in Fig. 1 denotes the backoff stage beyond which the contention window distribution remains unchanged (i.e., $p_l^{(m+i)} = p_l^{(m)}, \forall i, l \geq 0$) and it can be taken to be equal to infinity. Note that in the standard [24] the station draws a backoff counter value uniformly in the range from 0 to $W_i - 1$, so $p_l^{(i)} = 1/W_i$, $0 \leq l \leq W_i - 1$, $\forall i \geq 0$ and the upper margin of the backoff window expands exponentially up to the threshold stage m (i.e., $W_i = 2^i W_0$ for $0 \leq i \leq m$ and $W_i = W_m$ for $i > m$).

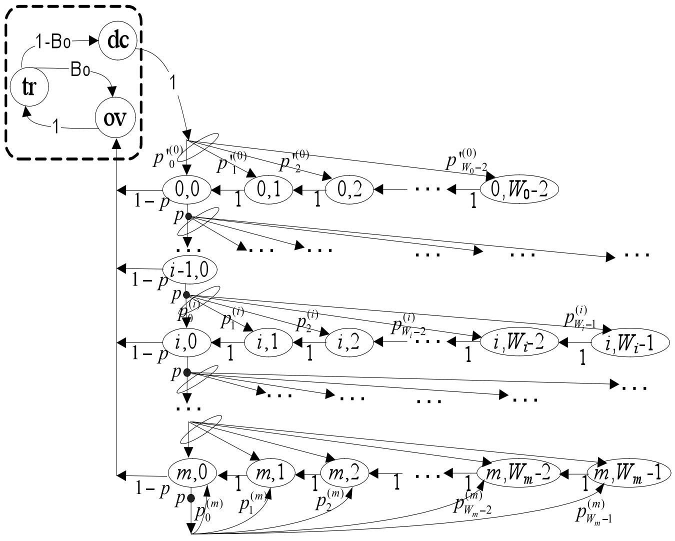

Figure 1: Discrete-time Markov chain for the backoff dynamics of IEEE 802.11 stations.

A station that has just completed a successful transmission is the only one that can immediately regain access to the wireless channel, because the other listening stations must wait for the passing of an empty system-slot to decrease their backoff counters before having the opportunity to transmit [4]. Such repeated access occurs with probability $B_0 \triangleq p_0^{(0)}$, the probability that a backoff counter sampled at the 0th stage has zero value. Consequently, the total successful transmission time seen by the listening stations is equal to the sum of the durations of the consecutive successful transmissions (whose number is geometrically distributed with parameter B_0) plus the duration of an empty system-slot needed for the backoff counters decrement.

It follows that, in order to represent properly all the consecutive successful transmissions after one visit at state $(i, 0)$, $i \geq 0$, and the following empty system-slot, the transition probabilities drawn at the 0th stage, $p_l^{\prime(0)}$ are different that $p_l^{(0)}$. Indeed, when exiting from the sequence of successful transmissions, the transmitting station draws a nonzero backoff counter $b \geq 1$. Afterwards, it follows necessarily an empty system-slot, at the end of which all stations have decreased their backoff counter and become synchronized. Since the backoff counter b will be observed/used after this initial decrement/synchronization, the observed value will appear to be one unit smaller. Therefore, the probability $p_l^{\prime(0)}$ that the (observed) value of the backoff counter b at the 0th stage is l is equal to

$$\begin{aligned} p_l^{\prime(0)} &= \Pr\{b \text{ at } 0^{\text{th}} \text{ stage} = l + 1 \,|\, b > 0\}, \\ &= \frac{p_{l+1}^{(0)}}{1 - B_0}, \quad \forall l \geq 0. \end{aligned} \tag{16}$$

From the previous discussion it is clear that the sampling of zero backoff counter values involved between successive successful transitions of a station do not correspond to transitions in the Markov chain of Fig. 1.

Solving the Markov chain of Fig. 1 for the steady-state probabilities $b_{i,j}$ by a methodology completely analogous to the one used for the slightly simpler Markov chain of [2], one obtains the transmission probability

$$\tau = \sum_{i=0}^{\infty} b_{i,0} = \left[1 + (1 - p)\Big(\frac{\overline{W}_0}{1 - B_0} - 1 + \sum_{i=1}^{\infty} p^i \overline{W}_i\Big)\right]^{-1}, \tag{17}$$

where $\overline{W}_i \triangleq \sum_{j=0}^{\infty} j p_j^{(i)}$ is the mean backoff window at the i^{th} stage, $i \geq 0$. The expression $\overline{W}_0/(1 - B_0) - 1$ is the adjusted mean backoff window at the 0th stage, corresponding to the probabilities in (16). If there is a finite threshold stage m beyond which the backoff window distribution remains unchanged (as in the standard), (17) specializes to

$$\tau = \left[1 + (1-p)\Big(\frac{\overline{W}_0}{1-B_0} - 1 + \sum_{i=1}^{m-1} p^i \overline{W}_i + \frac{p^m \overline{W}_m}{1-p} \Big) \right]^{-1} .$$

Since the Markov chains of the different wireless stations evolve independently from each other, one also obtains

$$1 - p = (1-\tau)^{n-1}, \tag{18}$$

because a station emitting a packet will not suffer a collision exactly when none of the other $n-1$ stations attempt to transmit. Equations (17) and (18) provide relations between p and τ that can be solved uniquely [2] for the values of these parameters. The values of p and τ depend on the number of competing stations n, due to (18).

At this point it is remarked that the difference of the backoff window distribution at stage-0 (i.e., the usage of $p_l'^{(0)}$ in (16) instead of the original probabilities $p_l^{(0)}$) and the allowance for arbitrary counter probability distribution functions $p_l^{(i)}$ constitute the differences between the Markov chain in [2] and the one discussed here and in Fig. 1. The Markov chain model adopted here is entirely equivalent to the setting captured in [4], where the analysis is carried out by means of renewal theory arguments, without direct reference to a Markov chain setting.

As already noted, the preceding analysis assumes that all stations are saturated. This paper employs the values of p and τ, obtained from (17) and (18), to calculate the Eff. Capacity of an IEEE 802.11 station, effectively assuming that all other stations are saturated. This approximation is on the safe side (i.e., 'conservative') since assuming the other stations saturated corresponds to the worst case. Section 5 will discuss how the saturation assumption can be waived and the model be applicable to all network load settings.

4.2. Semi-Markovian Characterization and Effective Capacity Derivation

4.2.1. The Basic Semi-Markovian Model

As already mentioned, the model of Fig. 1 (not including the states in the dashed box) is a discrete Markov chain, describing a process embedded at the instances when the listening stations decrease their backoff counters. At these instances the processes corresponding to different stations become synchronized (not by an arbitrary model assumption, but by virtue of the IEEE 802.11 MAC protocol). This model is exactly what is needed for computing the transmission probability τ as a function of the conditional collision probability p, because an attempt to transmit is equivalent to entrance into any one of the states $(i, 0)$. Note that this model does not include details related to actual transmission; the latter just occurs in between the transition from some state $(i, 0)$ to some state $(0, j)$.

Now that p and τ have been calculated through (17) and (18), i.e, on the basis of the embedded process, one may proceed to incorporate the details of packet transmission by expanding the Markov chain to include the three additional states in the dashed box, labelled `ov`, `tr`, and `dc`. State `ov` corresponds to the signaling/overhead transmissions before and after a data packet transmission, plus the transmission dedicated to the packet's MAC header. State `tr` corresponds to the successful transmission of the packet's payload, while State `dc` to the constant shortest system time-slot, required for the initial decrement of the backoff counter before the backoff is entered after a successful transmission [4]. Note that, by virtue of the IEEE 802.11 MAC, only one of the competing stations may reside in one of the States `ov`, `tr`, `dc`; all others must necessarily be in some original backoff state each.

With respect to the transition probabilities, the overhead and payload data transmissions always occur in pairs, thus the transitions from State `ov` to State `tr` occur with probability one. After a successful payload transmission, State `ov` is visited again with probability B_0 (probability that the backoff counter value at the 0th stage is zero) so the station transmits successfully one more time, while with probability $1 - B_0$ the station enters the backoff procedure after the backoff counter has been decreased by one (State `dc`).

Furthermore, one may associate the evolution of the expanded discrete model in Fig. 1 with the actual passage of time, by assigning the proper sojourn times to all involved states. These sojourn times are now described in terms of the respective moment generators.

The sojourn time in State `ov` is constant, with moment generator

$$\gamma_{\text{ov}}(\omega) = e^{\omega t_{\text{ov}}}, \tag{19}$$

where

$$t_{\text{ov}} \triangleq (RTS + CTS + PHY_{\text{hdr}} + ACK)/r_{\text{signal}} + MAC_{\text{hdr}}/\hat{r} + 3SIFS + DIFS \tag{20}$$

is the overhead time preceding and succeeding the payload transmission. The quantities RTS, CTS, $SIFS$, $DIFS$, ACK, MAC_{hdr}, PHY_{hdr} denote respectively the RTS packet size, the CTS packet size, the SIFS time interval, the DIFS time interval, the ACK packet size, the MAC header and the PHY header. The quantity $\hat{r}$ is the nominal bit rate of the IEEE 802.11 shared channel, while r_{signal} is the transmission rate used for signaling operations. For compatibility with early versions of the IEEE 802.11 protocol, r_{signal} is equal to the nominal channel rate used in the original IEEE 802.11 version, i.e., $r_{\text{signal}} = 1$ Mbps. In more recent versions, $\hat{r} > r_{\text{signal}}$. The values of all these parameters are determined by the IEEE 802.11 standard [24–27].

The moment generator of the sojourn time in State `tr` (i.e., payload transmission time) is

$$\gamma_{\text{tr}}(\omega) = \mathrm{E}\left[e^{\omega P/\hat{r}}\right], \tag{21}$$

where P denotes the payload size. If P is constant, the transmission time is deterministic.

The moment generator of the sojourn time in State `dc` is again deterministic, with

$$\gamma_{\text{dc}}(\omega) = e^{\omega t_{\text{slot}}}, \tag{22}$$

where t_{slot} stands for the duration of the shortest system-slot (required for the initial backoff counter decrement).

Finally, the random sojourn time associated with each of the original states (all, except `ov`, `tr`, `dc`) is the random time required for decreasing the value of the backoff counter by one. By the initial model assumptions (i.e., independent evolution of Markov chains for different stations, invariant conditional collision probability p), these sojourn times are independent and identically distributed as a generic random variable T_s, which can take three distinct values: (a) the duration of a collision plus the duration of the shortest system-slot; (b) the duration of a successful transmission plus the duration of the shortest system-slot; (c) the duration of the shortest system-slot. The shortest system-slot included in each possible duration is necessary for the decrement of the backoff counter of the listening stations. Furthermore, the deterministic duration of a collision t_{coll} is equal to

$$t_{\text{coll}} \triangleq RTS/r_{\text{signal}} + EIFS + t_{\text{slot}}, \tag{23}$$

because collided RTS messages are followed by an $EIFS$ time interval plus the duration of the shortest system-slot. Based on the above,

$$\begin{aligned}\gamma_s(\omega) &\triangleq \mathrm{E}\left[e^{\omega T_s}\right] \\ &= P_{\text{coll}} e^{\omega t_{\text{coll}}} + P_{\text{empty}} e^{\omega t_{\text{slot}}} \\ &\quad + P_{\text{succ}} \frac{(1 - B_0)\gamma_{\text{tr}}(\omega) e^{\omega t_{\text{ov}}}}{1 - B_0 \gamma_{\text{tr}}(\omega) e^{\omega t_{\text{ov}}}} e^{\omega t_{\text{slot}}},\end{aligned} \tag{24}$$

where

$$\begin{aligned}P_{\text{succ}} &= (n-1)\tau(1-\tau)^{n-2}, \\ P_{\text{empty}} &= (1-\tau)^{n-1}, \\ P_{\text{coll}} &= 1 - P_{\text{succ}} - P_{\text{empty}},\end{aligned} \tag{25}$$

are the probabilities of a successful transmission, an empty slot and a collision respectively, observed by a station backing-off (which observes $n-1$ other independent stations). The last term in (24) expresses the fact that the total duration of a successful transmission results from a geometric series of individual transmissions by the station not backing-off, as previously explained.

In recapitulation, by associating the discrete-time Markov chain of Fig. 1 (with its state-space augmented to include States `ov`, `tr`, `dc`) with the sojourn times just described, one obtains a corresponding Semi-Markovian model. To

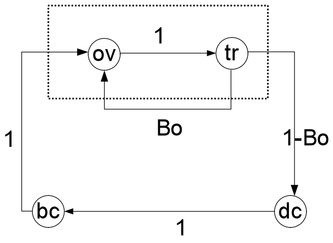

Figure 2: Equivalent Semi-Markovian chain with state-space compaction.

further reinforce the correspondence between this Semi-Markovian model and the original discrete-time model, one can start with the Semi-Markovian process and consider an embedding at instances where listening stations decrease their backoff counter. Then, the states occupied at these instances are exactly the original states (i, j), i.e., the diagram of Fig. 1 without the dashed box. It is noted that the Semi-Markovian model is essentially applied (although not explicitly mentioned) in both [2] (using a chain slightly simpler than that of Fig. 1) and [4], in the derivation of the saturation throughput.

The Semi-Markovian model is constructed in a way such that during the sojourn at each of its states a constant server rate is maintained. This rate is $\hat{r}$ (the WLAN's nominal peak rate) for State `tr` and zero for all other states. For the State `ov`, corresponding to signaling/overhead before and after payload transmission, including packet header transmission, we assign a service rate equal to zero even though the station actually transmits signaling data and/or the packet header. This choice is in line with the intention of studying the station's server rate available to the higher layers of the protocol stack.

4.2.2. State Compaction

Using the Semi-Markovian model just described, one may in principle directly apply the results of Section 3 towards obtaining the Eff. Capacity function of an IEEE 802.11 WLAN station. However, the state-space of this Markov chain is large (possibly infinite, if the threshold m or the upper bound of any backoff window distribution is infinite), making the direct application just mentioned computationally complex, if not intractable. Fortunately, since all the states corresponding to the backoff mechanism have zero service rate, they may be compacted in a single State `bc`, using a modified sojourn time T_{bc}, which is the random time between a departure from State `dc` and the subsequent entrance to State `ov` (see Fig. 1).

With this state compaction, one obtains the reduced four-states Semi-Markovian model of Fig. 2. State `bc` is the compacted state with sojourn time T_{bc}. As already explained, the service rate in State `tr` is equal to the nominal channel bit rate $\hat{r}$, while in all other states the service rate is zero. Clearly, the service rate process of the original Semi-Markovian chain in Fig. 1 is stochastically equivalent to the service rate process of the Semi-Markovian model in Fig. 2.

The sojourn time T_{bc} in State `bc` is described in terms of the moment generator $\gamma_{\mathrm{bc}}(\omega)$, which reflects the backoff dynamics as governed by the more detailed Semi-Markovian chain of Fig. 1. Towards deriving $\gamma_{\mathrm{bc}}(\omega)$, it is recalled that the number of collisions l during a backoff is geometrically distributed with parameter p. Conditioned on the fact that the number of backoff stages before each successful transmission is $l+1$ (resulting from l collisions and the final successful transmission) the backoff time is the sum of $l+1$ independent times spent in the $l+1$ backoff stages, plus l times spent in unsuccessful transmissions leading to collisions (each one after a backoff stage except the last one). Thus,

$$\gamma_{\mathrm{bc}}(\omega) = \sum_{l=0}^{\infty}(1-p)p^l\Big[(e^{\omega t_{\mathrm{coll}}})^l \hat{\gamma}_b^{(0)}(\omega)\prod_{j=1}^{l}\gamma_b^{(j)}(\omega)\Big], \tag{26}$$

where $\gamma_b^{(j)}(\omega)$ denotes the moment generator of the time spent at the j^{th} backoff stage. The function $\hat{\gamma}_b^{(0)}(\omega)$ is the moment generator of the time spent at the 0^{th} backoff stage, taking into account that the backoff counter value is nonzero and has been initially decreased by one, so $\hat{\gamma}_b^{(0)}(\omega)$ is different from $\gamma_b^{(0)}(\omega)$. Setting $T_b^{(j)}$ for the random time spent at the j^{th} backoff stage and b_j for the backoff window drawn at this stage, one has

$$\begin{aligned}
\gamma_b^{(j)}(\omega) &= \mathrm{E}\left[e^{\omega T_b^{(j)}}\right] = \mathrm{E}\left[\mathrm{E}\left[e^{\omega T_b^{(j)}} \mid b_j\right]\right] = \mathrm{E}\left[\gamma_s^{b_j}(\omega)\right] \\
&= \sum_{l=0}^{\infty} p_l^{(j)} \gamma_s^l(\omega) = g_j(\gamma_s(\omega)), \quad \forall j \geq 1,
\end{aligned} \tag{27}$$

where $\gamma_s(\omega)$ is as in (24) and where $g_j(z)$, $j \geq 0$, is the probability generator function of the backoff window distribution associated with the j^{th} stage, namely, $g_j(z) \triangleq \sum_{l=0}^{\infty} p_l^{(j)} z^l$. By definition, $g_0(0) = B_0$ and $g_j'(1) = \overline{W}_j$, $\forall j \geq 0$. Furthermore, in view of (16),

$$\hat{\gamma}_b^{(0)}(\omega) = \sum_{l=0}^{\infty} p_l'^{(0)} \gamma_s^l(\omega) = \frac{g_0(\gamma_s(\omega)) - B_0}{(1 - B_0)\gamma_s(\omega)}, \tag{28}$$

where $g_0(\omega)$ is the probability generator function of the initial backoff window distribution at the 0^{th} stage, in which zero backoff values occur with probability $p_0^{(0)} = g_0(0) = B_0$.

Combining (26), (27), and (28), we obtain

$$\begin{aligned}
\gamma_{\text{bc}}(\omega) = &\frac{g_0(\gamma_s(\omega)) - B_0}{\gamma_s(\omega)(1 - B_0)} \\
&\times \sum_{l=0}^{\infty} \Big((1-p)p^l e^{l\omega t_{\text{coll}}} \prod_{j=1}^{l} g_j(\gamma_s(\omega))\Big).
\end{aligned} \tag{29}$$

For the uniform backoff window distribution described in the standard,

$$g_j(z) = \sum_{l=0}^{W_j - 1} \frac{1}{W_j} z^l = \frac{1}{W_j} \frac{z^{W_j} - 1}{z - 1},$$

where $W_j = 2^{\min\{j,m\}} W_0$, for $j \geq 0$. If m is finite, as in the standard, then (29) specializes to

$$\begin{aligned}
\gamma_{\text{bc}}(\omega) = &\frac{g_0(\gamma_s(\omega)) - B_0}{\gamma_s(\omega)(1 - B_0)} \\
&\times \left[\sum_{l=0}^{m-1} \Big((1-p)p^l e^{l\omega t_{\text{coll}}} \prod_{j=1}^{l} g_j(\gamma_s(\omega))\Big) \right. \\
&\left. + \frac{(1-p)(p e^{\omega t_{\text{coll}}})^m \prod_{j=1}^{m} g_j(\gamma_s(\omega))}{1 - p g_m(\gamma_s(\omega)) e^{\omega t_{\text{coll}}}} \right],
\end{aligned}$$

with effective domain $\Omega_{\text{bc}} = (-\infty, \omega_{\text{bc}}^*)$, where ω_{bc}^* is the unique positive solution of $p g_m(\gamma_s(\omega)) e^{\omega t_{\text{coll}}} = 1$. In general, the upper bound ω_{bc}^* of the effective domain Ω_{bc} is determined by the requirement that the infinite sum in (29) converges. A sufficient (but not necessary) condition to have $\omega_{\text{bc}}^* > 0$ is that the backoff counter values drawn at stage j satisfy a boundedness condition as $j \to \infty$, which always holds in practical systems (as well as whenever $m < \infty$). Moreover, there is a corresponding necessary condition (again, always satisfied in practical systems) for $\omega_{\text{bc}}^* > 0$, described in Appendix A. There is no issue with the other States `ov`, `tr`, `dc` because, as it can be seen from (19), (21) and (22), the corresponding effective domains span the entire set of real numbers.

We conclude the discussion about the time spent by the station in backoff mode, by mentioning a fine point that was glossed over in Subsection 4.2.1 for the sake of maintaining clarity: While the sojourn times in states (i, j), $j > 0$ are the times required for backoff counter decrement, the sojourn times at states $(i, 0)$ are zero. This is exactly what is required when a successful transmission follows (i.e., a transition from $(i, 0)$ to State `ov` occurs, with probability $1-p$). When $(i, 0)$ is followed by a collision (i.e., a transition from $(i, 0)$ to $(i+1, j)$ occurs, with probability $p p_j^{(i+1)}$), time equal to t_{coll} passes before entering $(i+1, j)$. Thus, to be pedantic, we should expand the state diagram of Fig. 1 to include states in between $(i, 0)$ and $(i+1, j)$, each with a deterministic sojourn time equal to t_{coll} (and a zero service rate, of course). Actually, it is with the introduction of these states that the correspondence between the Semi-Markovian model and the original discrete embedded model is made complete. As it can be easily checked, the moment generator of the 'compacted' state in (29) accounts properly for this.

4.2.3. On/Off Representation and Effective Capacity Derivation

Another reduction can be imposed on the four-states model of Fig. 2, by recognizing that it is equivalent to a server model of the On/Off type. Indeed, State `tr` is the On state (with rate $\hat{r}$) and sojourn time moment generator

$$\gamma_{\text{on}}(\omega) = \gamma_{\text{tr}}(\omega) = \mathrm{E}\left[e^{\omega P/\hat{r}}\right]. \tag{30}$$

Upon departure from this state, the model enters the Off period: with probability B_0 its duration is equal to the sojourn time of State `ov`, while with probability $1 - B_0$ it is equal to the sum of the sojourn times in States `dc`, `bc`, and `ov`. Therefore, the moment generator of the sojourn time for the Off period is

$$\gamma_{\text{off}}(\omega) = \gamma_{\text{ov}}(\omega)\Big(B_0 + (1 - B_0)\gamma_{\text{bc}}(\omega)\gamma_{\text{dc}}(\omega)\Big), \tag{31}$$

where $\gamma_{\text{ov}}(\omega)$, $\gamma_{\text{dc}}(\omega)$, $\gamma_{\text{bc}}(\omega)$ are as in (19), (22), and (29). After either one of the two branches relevant to the Off period is followed, the On state is re-entered, as it should for an On/Off model. Using this alternative On/Off representation, the Eff. Capacity function can be derived from (10) where $u(s)$ is the unique negative solution of (14), in which $\gamma_{\text{on}}(\cdot)$ and $\gamma_{\text{off}}(\cdot)$ are as in (30) and (31).

It is noteworthy that, while the probabilities p and τ (needed as inputs to the model) depend only on the mean backoff values $\overline{W}_i$, $i \geq 0$, and the probability B_0 (through(17) and (18)), the full Eff. Capacity function relates to higher order characteristics of the backoff counter distributions. This observation is valuable for works trying to optimize the network performance through making the backoff parameters adaptive to the number of competing stations [29–31]. For a given number of competing stations, the network performance can be enhanced by maximizing the Eff. Capacity function through an appropriate selection of backoff counter distributions and this is the reason why general such distributions are used in the analysis.

The Eff. Capacity function derived in this section resulted from a consideration of the Semi-Markovian model in Fig. 2, which was shown stochastically equivalent (with respect to rate fluctuations) to the original Semi-Markovian model based on the Markov chain of Fig. 1, itself referring to the setting studied in [4]. Since by the general Eff. Capacity theory $a_C(0)$ is equal to the mean service rate, in this case equal to the saturation throughput, one has the following result:

Proposition 2. *The value $a_C(0)$ is equal to the station's saturation throughput, as derived in [4].*

The model equivalence just described suffices to establish Proposition 2. However, in the interest of further model verification we also prove it by independent analytical reasoning in Appendix B.

4.3. Effective Capacity-Based Traffic Control

The Eff. Capacity function of an IEEE 802.11 station can be employed in admission control tests, as explained at the end of Section 2. For an example, consider a scenario where a station already running a service needs to decide if it can engage another one, while maintaining the buffer overflow probability below some target threshold $e^{-\epsilon}$, given a buffer of size x. Assume that the preexisting service generates traffic at a constant bit rate $\mathcal{R}_{\text{cbr}}$ and that the second service features Poisson traffic with a constant packet size D and a Poisson rate λ (these parameters giving rise to a mean bit rate equal to $\mathcal{R}_{\text{pois}} = \lambda D$). Due to the additivity property of the Eff. Bandwidth functions, the Eff. Bandwidth of the aggregate traffic is $a_B(\theta) = \mathcal{R}_{\text{cbr}} + \lambda(e^{\theta D} - 1)/\theta$. The second service will be admissible if the inequality to the left of (8) holds for the Eff. Bandwidth function of the aggregate traffic just mentioned and $\theta = \epsilon/x$. It is reminded that, by virtue of Proposition 1, (8) is equivalent to (15). The simplicity of the admission control test mainly stems from the fact that the Eff. Capacity function of the observed station can be calculated without regard to the traffic load of other stations in the WLAN.

The same procedure applies to other input traffic profiles since, as already remarked, arbitrary traffic patterns may be handled, as long as the traffic possesses a well-defined Eff. Bandwidth function. For example, a bursty On/Off traffic profile modelled as a Markov Modulated Poisson Process (MMPP) with two states, mean On and Off durations equal to $1/\beta$ and $1/\alpha$, respectively, and Poisson arrival of packets (of constant size D) at rate λ_{MMPP} during the On period corresponds to the Eff. Bandwidth function

$$a_{\text{B,MMPP}}(\theta) = \frac{(e^{D\theta}-1)\lambda_{\text{MMPP}} - (\alpha+\beta) + \sqrt{((e^{D\theta}-1)\lambda_{\text{MMPP}} - (\alpha+\beta))^2 + 4\alpha(e^{D\theta}-1)\lambda_{\text{MMPP}}}}{2\theta}$$

(see, e.g., [32]). The mean traffic rate for this model is $a_{\text{B,MMPP}}(0) = \lambda_{\text{MMPP}} D\alpha/(\alpha+\beta)$.

4.4. Adaptation of the Model for the Basic Access Mode

Until now, it was assumed that the WLAN operated in the RTS/CTS access mode. However, the particular access mode in use affects the model only through the values of the overhead time t_{ov} and the collision time t_{coll} (see (20) and (23)). Specifically, in the Basic Access mode equation (20) must be replaced by

$$t_{\text{ov}} \triangleq (PHY_{\text{hdr}} + ACK)/r_{\text{signal}} + MAC_{\text{hdr}}/\hat{r} + SIFS + DIFS, \tag{32}$$

due to the changed signaling mechanism (i.e., direct data packet transmission instead of the RTS/CTS signaling). Similarly, (23) must change to

$$t_{\text{coll}} \triangleq PHY_{\text{hdr}}/r_{\text{signal}} + (MAC_{\text{hdr}} + P)/\hat{r} + DIFS + t_{\text{slot}}, \tag{33}$$

because the collided packets must now be fully transmitted before the collision is resolved. Note that, if the payload sizes of collided packets are variable in length, the longest such payload size must be used in (33).

Besides the two changes mentioned, all other expressions relevant to the Eff. Capacity model do not depend on the access mechanism and thus remain unchanged.

5. Model Adaptation to Non-Saturated Environments

The Eff. Capacity model developed in Subsection 4.2 relies on the saturation assumption and thus works well only when all other stations in the WLAN are heavily loaded. However, the dependence on the saturation condition is only through a few model parameters, namely the conditional collision probability p and the probabilities P_{succ}, P_{empty} and P_{coll} used in (24). Under non-saturation conditions these parameters retain their meaning, but take different values. If the correct values are supplied, the model works well in all settings, lightly loaded ones included.

To see this, consider a non-saturated station. Eventually, the station will have a zero backlog and, instead of trying to transmit, it will enter an "idle state" and remain there until a new packet enters its queue. Therefore, for all non-saturated stations, the extended state-space diagram in Fig. 1 must change to incorporate this "idle state".

However, it is important to note that this model change *does not* apply to the Markov chain of the 'tagged' station whose Eff. Capacity is to be determined. Indeed, for any kind of server, not just the one of the WLAN station, the Eff. Capacity is related (through (6)) to the stochastic behaviour of the data that *can be processed* in a time window $(0, t]$, asymptotically as $t \to \infty$. Thus the server must be studied as if it always had something to process, i.e., as if the corresponding station was saturated; in other words, the server's potential, rather than the actual output process, is of interest. (As an example, consider a server of constant rate c; its Eff. Capacity is equal to $a_C(\theta) = c$ for all arguments θ, regardless of the fact that whenever the backlog is zero the actual service rate is also zero, instead of c.)

Although the structure of the Markov chain of Fig. 1 corresponding to the tagged station remains unaffected, the chains of all unsaturated competing stations do change, to incorporate the "idle state". As a result, the WLAN is not homogeneous any longer, and there aren't common values of the parameters p and τ applicable to all stations. Rather, each station $i = 1, \ldots, n$ has its 'own' p_i and τ_i, connected by a relation of the form $\tau_i = f_i(p_i)$. For the tagged station, as well as all saturated stations that may exist in the WLAN, this relation continues to be expressed by (17). For each unsaturated station the functional form of $f_i(\cdot)$ is different, in order to incorporate the presence of the "idle state", including the value of the probability with which this state is occupied (expressing indirectly the station's traffic load). Moreover, (18) is now replaced by the n equations $1 - p_i = \prod_{j \neq i}(1 - \tau_j)$. Thus, there are $2n$ equations in total, which can be solved for the same number of unknowns p_i, τ_i, $i = 1, \ldots, n$. Once these are available, the quantities P_{succ},

P_{empty} and P_{coll} used in (24) may also be computed for station i, by heterogeneous analogs of (25) involving τ_j, $j \neq i$. (For example, the expression for P_{succ} becomes $P_{\text{succ}} = \sum_{j \neq i} \tau_j \prod_{k \neq j,i}(1 - \tau_k)$.)

Such a kind of 'heterogeneous' modeling for unsaturated environments has been successfully pursued in other works, e.g., [33, 34], primarily targeting throughput analysis. These works calculate the values p_i, τ_i by a methodology along the lines just described and then proceed to calculate throughput values much in the same way as in [2]. For our purposes, it is of interest to note that the success of the models in [33, 34] validates the key assumption that the conditional collision probabilities p_i remain invariant in the heterogeneous setting too, thus the state diagram of Fig. 1 still corresponds to a Markov chain. This property, along with the fact that the state diagram for the station whose Eff. Capacity is to be determined remains exactly as depicted in Fig. 1, enables repetition of the whole chain of reasoning within Subsection 4.2, thus establishing that the Eff. Capacity model remains of the same On/Off form as in the saturated environment; the only change is in the value of p in (29) and the values of P_{succ}, P_{empty} and P_{coll} in (24).

Therefore, if there is complete load knowledge of all unsaturated stations in the WLAN, one may calculate (by the methodology outlined above) the four parameters p, P_{succ}, P_{empty} and P_{coll} for the 'tagged' station and then feed these values in (29) and (24) to obtain the Eff. Capacity function of the station. However, this paper focuses on distributed admission control, without presupposing knowledge of traffic conditions at competing stations. Thus, it is suggested that the values of these four parameters be measured instead. Note that, since the 'tagged' station must be observed as if it was saturated, these probabilities must be sampled in a period during which the station transmits whenever it can access the channel, i.e., has positive backlog throughout the measurement period. For all modern versions of the IEEE 802.11 protocol, measurement periods in $O(1\,\text{s})$ suffice to ensure the reliable estimation of all values sought. In addition, note that, whenever the competing stations are close to saturation, the measured values of the four parameters will approach the values that would result from (17), (18) and (25). In conclusion, the Eff. Capacity model, using the modified values of p, P_{succ}, P_{empty} and P_{coll}, operates successfully in the whole range from saturated to lightly loaded environments.

It is remarked that, when the Eff. Capacity of an IEEE 802.11 station in an unsaturated environment is determined by the measurement-assisted methodology just described, the resulting Eff. Capacity function depends on the traffic load of all other competing stations (indirectly, through the values of the four probabilities measured by the station). In particular, the said Eff. Capacity function becomes smaller whenever the traffic load at any competing station increases (or whenever additional active stations enter the network). Keeping this in mind, consider an active IEEE 802.11 station wishing to admit a new traffic flow, subject to tail-related QoS constraints: According to the measurement-assisted process, the given station will measure the values of the four probabilities p, P_{succ}, P_{coll} and P_{empty} pertaining to its environment and will use them in (29) and (24), towards applying the admission control test (15) (i.e., the specialized form of the generic test (8)). The Eff. Bandwidth function $a_B(\cdot)$ used in the test (15) should include the new traffic flow tested for admission, in addition to the load of the flows already established through that station. The steps just described are adequate for preserving the QoS at the station admitting the new flow, but may fall short in a network-wide scope, because, as explained, once the new flow is admitted, the Eff. Capacities of the other stations will become smaller and the condition (15) may not hold anymore for the context relevant to some of them, signifying a QoS compromise therein.

It is possible to extend the methodology for ensuring QoS throughout the network, provided the stations are capable of exchanging some extra (non-standard) signaling and coordination messages, as now outlined: Once the station wishing to admit the new flow has performed the measurement and admission test described previously, and provided that the result of this test allows admission of the flow, this station should request (through a signaling message) from all other stations in the WLAN permission to add the flow. Then, each of these other stations, in sequence, should also measure the local values of the probabilities p, P_{succ}, P_{coll} and P_{empty} and use them for applying the test (15) in its own context. If this test succeeds, a message granting the permission is sent to the requesting station, otherwise the permission is denied (and the whole sequence of responses to the request is terminated).

During the time each station measures the local values of the four probabilities, both this station and the station having issued the request should attempt to transmit whenever they can access the channel, i.e., behave as being saturated[4]. The measuring station should do so in order to obtain a proper measurement, as explained earlier, while the requesting station should behave as saturated in order to compensate, in a conservative fashion, for the extra traffic

[4] This requirement provides the reason for having the stations reply to the requesting station in sequence, rather than in parallel, for otherwise the conditions during the measurement period would simulate a saturated WLAN.

packet payload P	8184 bits
MAC header	272 bits
PHY header	120 bits
ACK	112 bits + PHY header
RTS	160 bits + PHY header
CTS	112 bits + PHY header
Data Bit Rate $\hat{r}$	54 Mbit/s
Signaling Bit Rate r_{signal}	1 Mbit/s
Slot Time	$20\,\mu$s
SIFS	$10\,\mu$s
DIFS	$50\,\mu$s
initial backoff window W_0	32
maximum backoff stage m	5

Table 1: Parameters used in analytical and simulation results.

load that is about to admit. (One could envisage a more efficient alternative, according to which the requesting station would employ a traffic pattern representative of the traffic flow to be added, on top of the already established traffic, but the practical aspects of this alternative are harder to tackle.) The overall scheme is only mildly conservative, because, during each measurement, only one station behaves as if it was saturated.

6. Model Validation

We now validate the Eff. Capacity function of IEEE 802.11 stations by comparing analytical with simulation results. The simulation results were obtained with the help of the ns-2 simulator [35], using the system parameter values in Table 1. These values correspond to IEEE 802.11g WLAN, in Direct-Sequence Spread Spectrum (DSSS) – Orthogonal Frequency-Division Multiplexing (OFDM) short preamble mode [27]. All scenarios assume a constant payload size $P = 1023$ bytes, and the simulation results are presented with 95% confidence intervals.

Each line and associated points in Fig. 3 jointly correspond to the Eff. Capacity of a station within a network operating in the RTS/CTS mode and containing n stations in total. Different groups of lines/points correspond to different values of n. For a specific value of n, the line plots the respective Eff. Capacity function (more precisely, $a_C(-\theta)$ vs. θ), as derived by the saturation-based model. Each of the associated points encodes results from a corresponding simulation run, set up as now explained: One station is chosen for observation. Traffic of a known profile, thus also known Eff. Bandwidth function $a_B(\cdot)$, satisfying $r_{\min} < \hat{r}_{\text{input}}$ and $\bar{r} > \bar{r}_{\text{input}}$, is fed to the observed station, whose buffer is monitored. The quantities $\hat{r}_{\text{input}}$ and $\bar{r}_{\text{input}}$ denote the peak and mean traffic rates, respectively. All other stations feature a very high load, so as to become saturated.

For the given choice of input traffic, there is a value $\theta^\star$ that satisfies (3) and (7). This value $\theta^\star$ is determined by the simulation, which provides (at steady state) a histogram of the amount of data queued in the buffer, i.e., effectively provides the empirical Complementary Probability Distribution Function (CPDF) of the queue length, which is then converted to semi-log scale and the slope of its linear tail is determined (preferably through least-squares match, for increased precision). The negative of this slope is the estimated value of $\theta^\star$.

Due to (7), the unknown $a_C(-\theta^\star)$ is equal to the value $a_B(\theta^\star)$. This last quantity is known, since both the function $a_B(\cdot)$ and the parameter $\theta^\star$ are known (the second one from the simulation results). Thus, the simulation ultimately provides the associated point $(\theta^\star, a_C(-\theta^\star))$. Other such points are obtained by repeating this process (with different traffic—in terms of form, parameters, or both—otherwise the same point will result). Fig. 3 suggests a very close match of the simulation results to the model.

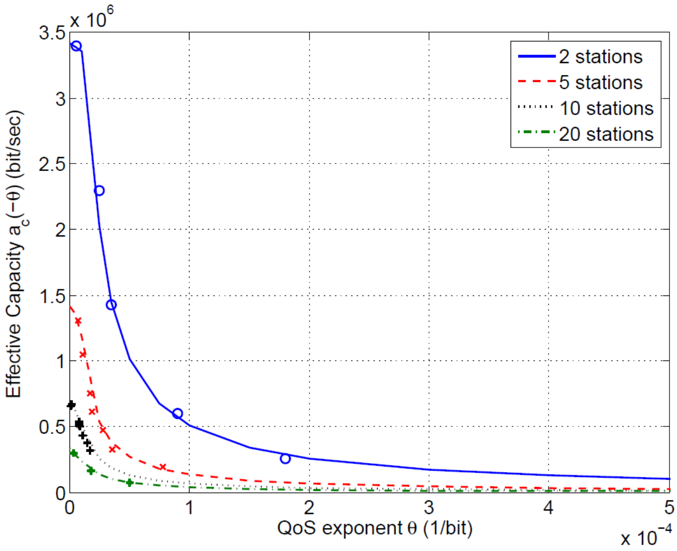

Figure 3: Effective Capacity function curves ($a_C(-\theta)$ vs. θ) for different network sizes.

We now turn into system performance evaluation in terms of the buffer overflow probability for a highly loaded environment. Figure 4 displays probabilities of exceeding queue thresholds for a station in a network operating in the RTS/CTS mode and containing 10 stations. The observed station is fed with: (a) an On/Off Markov Modulated Poisson Process (MMPP) (green lines), (b) Poisson traffic (red lines), and (c) cbr traffic (blue lines). In all of these cases, the same mean traffic rate of 650 kbps has been used. For traffic profile (a), the mean durations of the On and Off periods are both set equal to 1 s. The other stations in the WLAN are saturated. The decay rate θ^* of the tail-probabilities in the figure is analytically determined as $\theta^* = \sup\{\theta \mid a_B(\theta) \leq a_C(-\theta)\}$, this value being the unique positive solution of (7). As shown in Fig. 4, simulation-derived tail-probabilities ultimately decay exponentially and the decay rate agrees well with the one derived from analysis. Note that the model captures the dependence of the decay rate on the traffic details (expressed through the respective Eff. Bandwidth function). The decay rate corresponding to the cbr traffic is highest, while that of the MMPP traffic lowest, with the decay rate of the Poisson traffic in between the two extremes. This is in accordance with the different variability/burstiness present in each of these three traffic models (all of whom share the same mean traffic rate).

Fig. 5 provides results for a scenario with much less loaded competing stations, again for the RTS/CTS access mode. The station whose queue is observed is loaded with On/Off MMPP traffic of mean rate 1.2 Mbps. As previously, the mean On and Off durations are equal to 1 s. The other 9 competing stations feature a Poisson load of mean rate 500 kbps. Since the 'background' stations are away from the saturation regime, the Eff. Capacity of the observed station is determined by employing the adaptation of the model for non-saturated environments. As explained in Section 5, the four parameters p, P_{succ}, P_{empty} and P_{coll} must be measured. For this purpose, a brief simulation is initially run to assess the values of these parameters, which are then employed in (29) and (24). Apart from this step, the decay rate of the queue content tail-probabilities is analytically determined as for the saturation-based model. The dotted line in Fig. 5 plots the tail-probabilities as determined by simulation, while the solid straight line has slope equal to the analytically-derived decay rate, obtained by the methodology just described. It is seen that the model tracks correctly the asymptotic decay rate of the tail-probabilities in this case too.

Note that it would have been impossible to assess the asymptotic decay rate of the tail-probabilities on the basis of the saturation-based Eff. Capacity model. The saturation throughput for 10 competing stations is about 670.8 kbps and the mean traffic rate used equal to 1.2 Mbps, so the observed station would have appeared overloaded on the basis of that model and the resulting CPDF of the buffer occupancy would have degenerated to a horizontal straight line in semi-log scale. This fact provides evidence that the adaptation of the model described in Section 5 indeed goes beyond the original saturation-based one.

Towards further validation, the model's accuracy in non-saturated environments is now investigated directly in

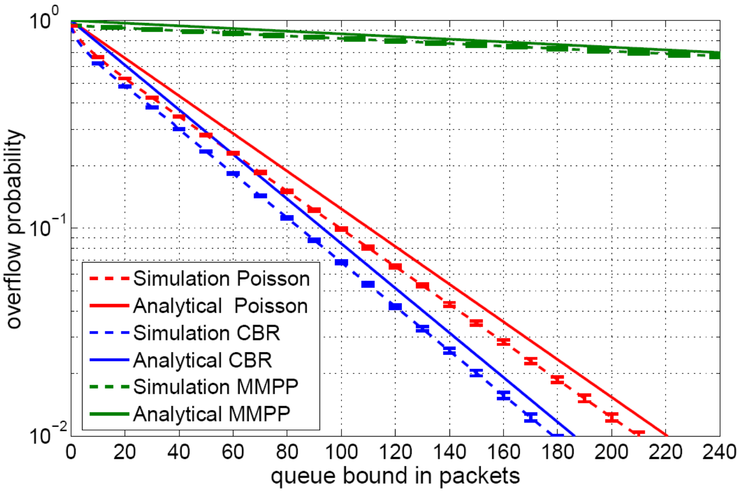

Figure 4: Analytical and simulation results (with 95% confidence intervals) for probabilities of exceeding queue-content thresholds. Observed station examined under three different traffic profiles with 9 other saturated stations.

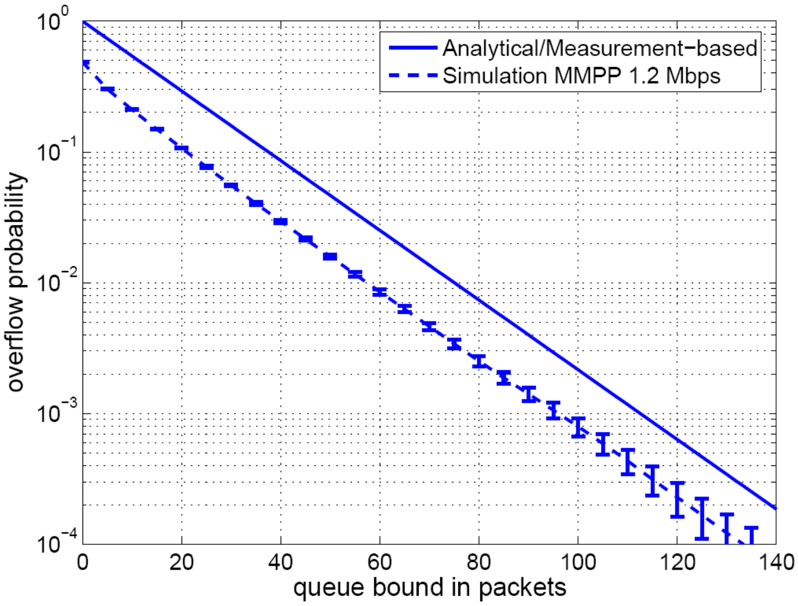

Figure 5: Analytical and simulation results (with 95% confidence intervals) for probabilities of exceeding queue-content thresholds. The observed station coexists with 9 other unsaturated stations.

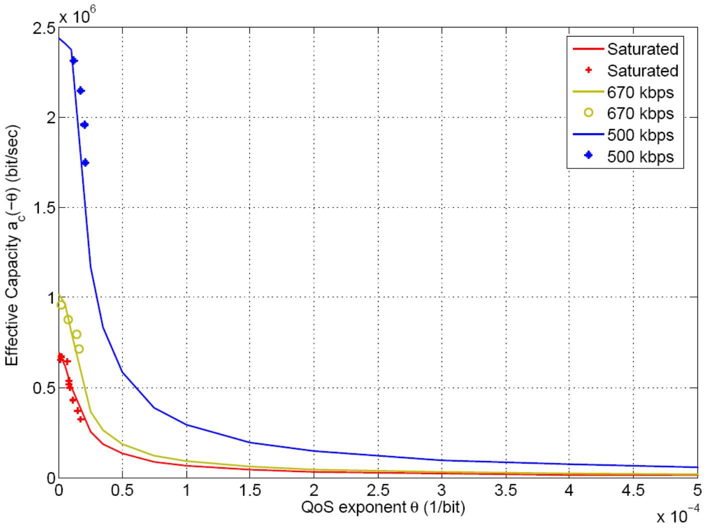

Figure 6: Eff. Capacity function curves ($a_C(-\theta)$ vs. θ) for a station in a WLAN with 9 additional stations featuring varying traffic loads.

terms of the Eff. Capacity function, in the spirit of the results for the saturation-based model within Fig. 3. Data displayed in Fig. 6 refer to the Eff. Capacity of a station in a WLAN operating in the RTS/CTS access mode and containing 9 additional competing stations. Different colors of lines and marks correspond to different loading conditions for the nine ‘background’ competing stations: Specifically, red corresponds to saturated background stations (and the relevant data had also been included in Fig. 3), while yellow and blue correspond to Poisson traffic loads of mean rate equal to 670 kbps and 500 kbps, respectively, for each of the background stations.

For any given color, the solid line is the model-derived Eff. Capacity function. Except for the saturated case (red), a very brief simulation run is again required to calculate the four input probabilities to the model. Note that a single brief simulation run suffices to determine the full Eff. Capacity function (the whole line of a given color in Fig. 6). Each mark of the same color corresponds to the assessment (by means of simulation), of a point of the form $(\theta, a_C(-\theta))$, by the same methodology as the one used for Fig. 3. Each point requires a simulation run comparable to the one used for obtaining Fig. 5 (or Fig. 4). Thus, Fig. 6 essentially encapsulates several different figures of the form of Fig. 5. Therefore, the results in Fig. 6 provide a strong indication that the model is successful in light load conditions as well.

We now focus on QoS admission control decisions based on the IEEE 802.11 Eff. Capacity function. The considered WLAN operates in the RTS/CTS access mode and contains 10 stations, all of which feature the same Poisson traffic, of a mean rate equal to 600 kbps. One of the stations attempts to initiate additional flows on top of the existing Poisson background traffic, one after the other. The traffic profile of each of these flows is of the (fluid, rather than MMPP) On/Off type, with exponentially distributed On and Off periods of mean durations equal to 0.4 s and 0.8 s, respectively, and a peak rate equal to 480 kbps. These parameters yield a mean rate of 160 kbps per flow.

Admission control is exercised to assess whether a flow may be admitted on top of the previously existing traffic without violating the QoS specification, which states that queue lengths greater than 120 packets should occur with probability at most 10^{-2}. According to the comments in Subsection 4.3, the value of θ to be used in the admission control test (8) (equivalently, (15)) is equal to $\theta = -\log 10^{-2}/(120 \times 1023 \times 8)\,\text{bit}^{-1}$. When testing for admission of the k^{th} On/Off flow, the Eff. Bandwidth function in (15) is set to $a_B(\cdot) = a_{\text{Poisson}}(\cdot) + k a_{\text{onoff}}(\cdot)$. Note that since the other competing stations are not saturated, the measured values of the conditional collision probability p and the probabilities P_{succ}, P_{empty} and P_{coll} are different from the values that would have been obtained in a saturated environment, leading to a different (greater) Eff. Capacity function.

For the scenario considered, the admission control procedure accepts up to 4 On/Off flows in addition to the Poisson background traffic. The correctness of this decision is validated by Fig. 7, which plots probabilities of exceeding queue-length thresholds when the station is loaded with 4 (blue lines) and 5 (red lines) On/Off flows in addition to the

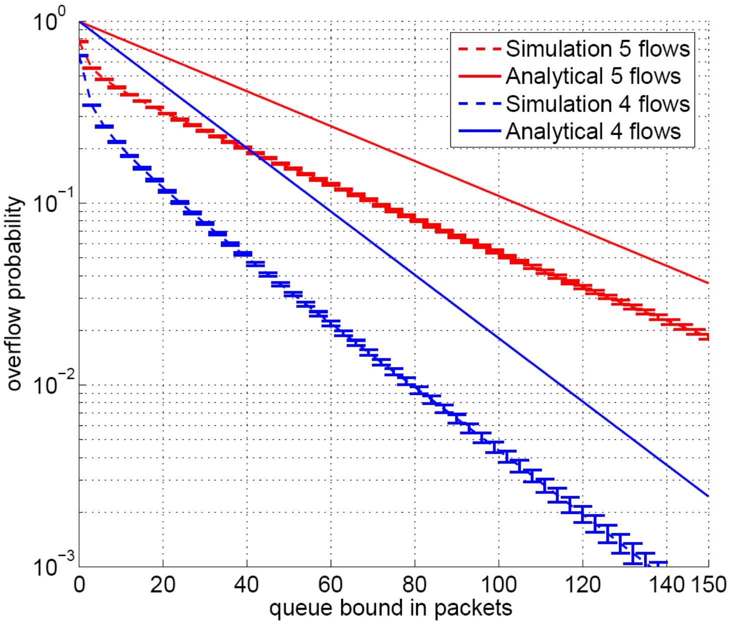

Figure 7: Queue-content tail-probabilities when the station admits four and five On/Off flows, on top of Poisson background load.

	Poisson	MMPP	Poisson + MMPP
Saturation-based AC	8	3	5
Measurement-assisted AC	9	6	8
Exact Max # of Stations	9	6	8

Table 2: Maximum number of IEEE 802.11 stations that can be admitted to the WLAN subject to a QoS constraint, for different traffic scenarios.

Poisson background load. Results from both simulation (dashed lines) and analysis (solid lines) are included. It may be seen that with 4 flows the probability of the queue length exceeding the threshold of 120 packets is below the target value 10^{-2} and that the introduction of the 5th flow raises the value of this probability above the threshold, violating the QoS.

To assess further the efficiency of the admission control scheme, this time in terms of the WLAN utilization, the next set of results discusses the maximum number of IEEE 802.11 stations that can be admitted to the WLAN subject to a given QoS constraint. More specifically, the WLAN is populated by an increasing number of IEEE 802.11 stations (all of whom generate traffic according to the same profile) until the QoS specification is violated. The exact maximum number of stations is determined by simulation and is compared against the number achieved when the admission control scheme is in effect (i.e., the sequence number of the last station for which the admission control test is successful). The results cover admission control based on both variants of the Eff. Capacity model for three different traffic profiles, namely: Poisson, On/Off MMPP (with mean On and Off periods equal to 0.5 s and 1 s, respectively), and a mixture of Poisson and On/Off MMPP (the two traffic components in this last case being of equal mean rates). All three traffic profiles yield the same mean rate, equal to 700 kbps. The QoS requirement in use is that the probability with which queues grow beyond 100 packets should not exceed 10^{-2}.

The results appear in Table 2. It may be observed that both variants of the admission control scheme guard against QoS violation. As expected, admission control based on the saturation assumption is somewhat conservative (compare rows 1 and 3 in the table). The measurement-assisted variant is more efficient and attains the maximum number of IEEE 802.11 stations that can be admitted, for all traffic scenarios (see rows 2 and 3 in the table). Note that, as the traffic becomes burstier, fewer stations can be admitted without compromising the QoS.

We close by validating the applicability of the Eff. Capacity model for WLANs using the Basic Access mode.

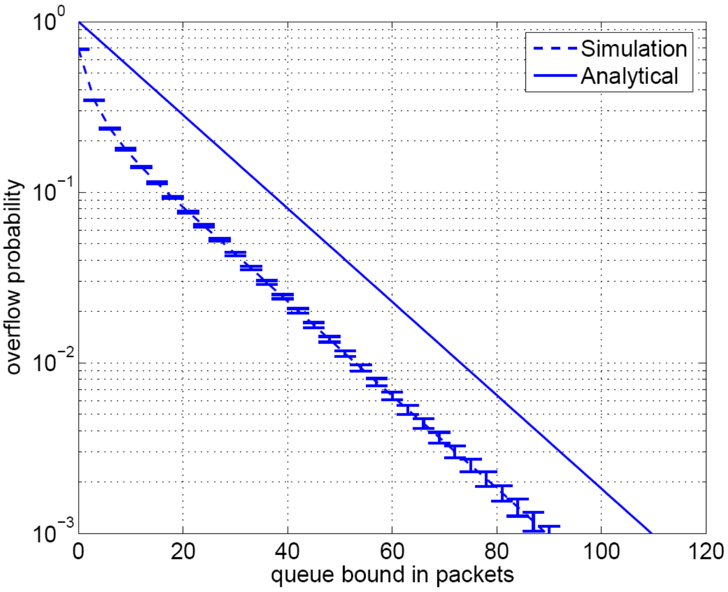

Figure 8: Analytical and simulation results (with 95% confidence intervals) for probabilities of exceeding queue-content thresholds for a station in a WLAN employing the Basic Access mode.

Analogously to the setting of Fig. 4, Fig. 8 displays probabilities of exceeding queue thresholds, but this time the Basic Access mode is in effect. The observed station is fed with Poisson traffic featuring a mean rate of 1.1 Mbps, while the other 9 stations in the WLAN are saturated. It can be seen that, as with the RTS/CTS access mode, the model tracks well the true asymptotic decay rate (determined by simulation) for WLANs with Basic Access mode too.

To highlight more clearly the effect of the access mode on the Eff. Capacity function, Fig. 9 displays the Eff. Capacity curves corresponding to the two access modes, for an observed station in a WLAN with 9 other saturated stations. As it can be observed, for the particular setting relevant to Fig. 9, the Eff. Capacity of the Basic Access mode is greater than its RTS/CTS counterpart, for all values of the parameter. This happens because the values of the overhead time t_{ov} and of the collision time t_{coll} in the Basic Access mode are both smaller than the respective values in the RTS/CTS access mode. These smaller values lead to a smaller generator function $\gamma_{\mathrm{off}}(\cdot)$ (see (31), (19), (29) and (24)) and finally to a greater Eff. Capacity function. The smaller value of t_{ov} in the Basic Access mode results from the reduced signaling (compare (32) with (20)). The smaller value of t_{coll} is due to the significant difference between the values of the signaling rate r_{signal} and the data rate $\hat{r}$ in IEEE 802.11g WLANs. Specifically, in the case of the RTS/CTS access mode, the collided RTS packets (transmitted with rate $r_{\mathrm{signal}} = 1$ Mbps—see (23)) last more than the time that would have been spent for the complete transmission of collided data packets in the Basic Access mode (where such packets are transmitted with rate $\hat{r} = 54$ Mbps—see (33)). The situation would have been reversed, had the payload size of packets been greater than $P = 2518$ bytes, in which case the collision time in the Basic Access mode would have been greater than its counterpart for the RTS/CTS access mode.

7. Conclusions

In this paper we developed an Eff. Capacity model suitable for use in IEEE 802.11 WLANs under DCF mode. Careful examination and appropriate application of previous modeling approaches and their combination with results from the Eff. Bandwidth/Capacity theory allowed the representation of an IEEE 802.11 station as a server of the On/Off type with well-defined characteristics. The Eff. Capacity associated with this model may be used, either for assessing tail-related performance of the station under a given load, or for applying traffic admission control tests to proactively enforce tail-related statistical QoS guarantees. In this second, and more important, application it is not even necessary to fully calculate the Eff. Capacity function associated with the On/Off model; an alternative,

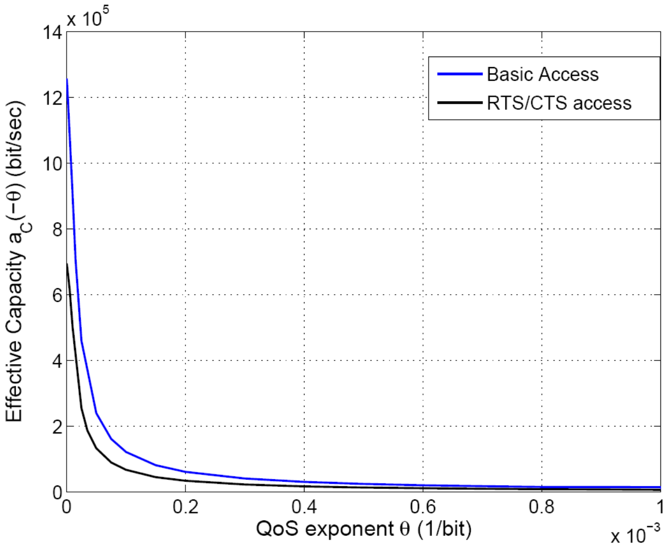

Figure 9: Eff. Capacity functions ($a_C(-\theta)$ vs. θ) for a station in a WLAN with 9 additional saturated stations and different access modes.

equivalent decision test may be applied, with virtually no computational cost. The model places no restriction on the type and complexity of the station's traffic load other than it has to possess a well defined Eff. Bandwidth function, which is the case for most processes of interest.

There are two variants of the model: a purely analytic, suitable for highly loaded WLANs, and a measurement-assisted, capable of coping with arbitrary networking environments. The second variant employs the measurement of a few simple parameters to indirectly assess network conditions. Only parameters related to the station's operation are measured; in particular no explicit knowledge of traffic conditions at other competing stations is required.

It is noted that, while the proposed model has been applied only to the DCF mode (ad hoc) in this paper, it can be readily adapted for use in environments with Access Points (infrastructure mode).

Appendix A. Necessary Condition for $\omega^*_{\mathrm{bc}} > 0$

Since $\omega^*_{\mathrm{bc}} > 0$, there exists $\omega > 0$ for which the infinite sum in (29) (equivalently (26)) converges. Then, by (27) and the convexity of the exponential function, $g_j(\gamma_s(\omega)) = \mathrm{E}\left[e^{\omega T_b^{(j)}}\right] \geq e^{\omega \mathrm{E}\left[T_b^{(j)}\right]} = e^{\omega \overline{W}_j \mathrm{E}[T_s]}$, where $\mathrm{E}\left[T_s\right] = \gamma'_s(0)$. As a result, the infinite sum in (26) converges for some $\omega > 0$ only if $\limsup_{l\to\infty}\left\{\log p + \omega t_{\mathrm{coll}} + \omega \mathrm{E}\left[T_s\right] \frac{1}{l}\sum_{j=1}^{l} \overline{W}_j\right\} \leq 0$, implying that the sequence of Cesàro sums $\frac{1}{l}\sum_{j=1}^{l}\overline{W}_j$ must remain bounded. This always holds in practical schemes and is automatically satisfied if the threshold stage m is finite. This condition will be used in the proof of Proposition 2.

Appendix B. Proof of Proposition 2

We first introduce the following lemma, used later on.

Lemma 1. *For arbitrary sequences a_j, b_j, $j \geq 0$:*

1. $\sum_{l=0}^{m} a_l \sum_{k=0}^{l} b_k = \sum_{l=0}^{m} b_l \sum_{k=l}^{m} a_k$, *for all* $m \geq 0$.

2. *If (i)* $\sum_{k=0}^{\infty} a_k$ *converges and (ii)* $\lim_{m\to\infty}\sum_{l=0}^{m} b_l \sum_{k=m+1}^{\infty} a_k = 0$, *then* $\sum_{l=0}^{\infty} a_l \sum_{k=0}^{l} b_k = \sum_{l=0}^{\infty} b_l \sum_{k=l}^{\infty} a_k$.

As a special case, the condition (ii) is satisfied if the Cesàro sums $\frac{1}{l+1}\sum_{j=0}^{l} b_j$ *are bounded and* $\lim_{n\to\infty} n \sum_{k=n}^{\infty} a_k = 0$.

Proof. Item 1 follows easily by induction. Using the result of this item, along with condition (i), one obtains

$$\sum_{l=0}^{m} a_l \sum_{k=0}^{l} b_k = \sum_{l=0}^{m} b_l \sum_{k=l}^{\infty} a_k - \sum_{l=0}^{m} b_l \sum_{k=m+1}^{\infty} a_k.$$

As m reaches infinity, Item 2 follows due to (ii).

For the special case, let M be a bound of $\left|\frac{1}{l+1}\sum_{j=0}^{l} b_j\right|$. Then $\left|\sum_{l=0}^{m} b_l \sum_{k=m+1}^{\infty} a_k\right| \leq M\left|(m+1)\sum_{k=m+1}^{\infty} a_k\right|$, and letting $m \to \infty$, the right hand-side of this last relation tends to 0. □

We now turn to the proof of the proposition. In view of (9), as specialized for On/Off models,

$$a_C(0) = \bar{r} = \hat{r}\mathrm{E}\left[T_{\text{on}}\right]/(\mathrm{E}\left[T_{\text{on}}\right] + \mathrm{E}\left[T_{\text{off}}\right]),$$

where $\mathrm{E}\left[T_i\right] = \gamma_i'(0)$, $i = \text{on}, \text{off}$. By (30), $\gamma_{\text{on}}'(0) = \mathrm{E}\left[P\right]/\hat{r}$; similarly (19), (22), and (31) lead to $\gamma_{\text{off}}'(0) = t_{\text{ov}} + (1 - B_0)(\mathrm{E}\left[T_{\text{bc}}\right] + t_{\text{slot}})$, with $\mathrm{E}\left[T_{\text{bc}}\right] = \gamma_{\text{bc}}'(0)$. Substitution yields

$$a_C(0) = \bar{r} = \frac{\mathrm{E}\left[P\right]/(1 - B_O)}{(\mathrm{E}\left[P\right]/\hat{r} + t_{\text{ov}})/(1 - B_0) + t_{\text{slot}} + \mathrm{E}\left[T_{\text{bc}}\right]}. \tag{B.1}$$

Consider (29) and denote by $\gamma_A(\omega)$ and $\gamma_B(\omega)$ the leading fraction and trailing sum therein, respectively. $\gamma_A(\cdot)$ is the moment generator of the time spent in the 0th backoff stage; similarly, $\gamma_B(\cdot)$ is the generator of the time collectively spent in all other stages (with index greater than 0) entered during the backoff interval. Clearly,

$$\mathrm{E}\left[T_{\text{bc}}\right] = \gamma_{\text{bc}}'(0) = \gamma_A'(0) + \gamma_B'(0). \tag{B.2}$$

By direct differentiation, also taking into account that $g_0'(1) = \overline{W}_0$, one obtains

$$\gamma_A'(0) = \left(\frac{\overline{W}_0}{1 - B_0} - 1\right)\mathrm{E}\left[T_s\right], \tag{B.3}$$

where

$$\begin{aligned} \mathrm{E}\left[T_s\right] = \gamma_s'(0) = &P_{\text{coll}}t_{\text{coll}} + P_{\text{empty}}t_{\text{slot}} \\ &+ P_{\text{succ}}\left(\frac{\mathrm{E}\left[P\right]/\hat{r} + t_{\text{ov}}}{1 - B_0} + t_{\text{slot}}\right), \end{aligned} \tag{B.4}$$

following from (24), is the mean time required for the reduction of the backoff counter by one.

In connection with $\gamma_B(\omega)$, let $S_l(\omega) \triangleq \prod_{j=1}^{l}\left(g_j(\gamma_s(\omega))e^{\omega t_{\text{coll}}}\right)$ be the moment generator of the time spent in the backoff stage with index $l \geq 1$. Then, $\gamma_B'(0) = \sum_{l=0}^{\infty}(1-p)p^l S_l'(0)$. Consideration of the associated cumulant generators $\phi_l(\omega) \triangleq \log S_l(\omega)$ leads to $\phi_l'(0) = S_l'(0) = \sum_{j=1}^{l}\left(\overline{W}_j\mathrm{E}\left[T_s\right] + t_{\text{coll}}\right)$ and substitution in the expression for $\gamma_B'(0)$ yields

$$\begin{aligned} \gamma_B'(0) &= \sum_{l=0}^{\infty}(1-p)p^l\sum_{j=1}^{l}\left(\overline{W}_j\mathrm{E}\left[T_s\right] + t_{\text{coll}}\right) \\ &= \sum_{j=1}^{\infty}\left(\overline{W}_j\mathrm{E}\left[T_s\right] + t_{\text{coll}}\right)\sum_{l=j}^{\infty}(1-p)p^l \\ &= \mathrm{E}\left[T_s\right]\sum_{j=1}^{\infty}p^j\overline{W}_j + \frac{p}{1-p}t_{\text{coll}}. \end{aligned} \tag{B.5}$$

The second equality in this last equation employs Lemma 1 (with $b_0 = 0$). This is possible because the sequence $\overline{W}_j$, as shown in Appendix A, is necessarily Cesàro sum bounded, $\sum_{l=0}^{\infty}(1-p)p^l$ converges and $\lim_{n\to\infty} n\sum_{k=n}^{\infty}(1-p)p^k = \lim_{n\to\infty} np^n = 0$.

By combining (B.2), (B.3), and (B.5), one obtains

$$\begin{aligned}\mathrm{E}\left[T_{\mathrm{bc}}\right] &= \frac{p}{1-p}t_{\mathrm{coll}} + \mathrm{E}\left[T_s\right]\left(\frac{\overline{W}_0}{1-B_0} - 1 + \sum_{l=1}^{\infty} p^l \overline{W}_l\right) \\ &= \frac{p}{1-p}t_{\mathrm{coll}} + \mathrm{E}\left[T_s\right]\frac{1-\tau}{\tau(1-p)}. \end{aligned} \tag{B.6}$$

The second equality employed (17) to replace the parenthesized expression in the first equality. It is noted that (17) is always valid, because, as explained in Section 5, when considering the Eff. Capacity of a station, this station always behaves as if it was saturated.

In order to express the results in the notation used by [4], we introduce the following equivalences:

$$\mathrm{E}\left[P'\right] \triangleq \frac{\mathrm{E}\left[P\right]}{1-B_0}, \quad T'_s \triangleq \frac{\mathrm{E}\left[P\right]/\hat{r} + t_{\mathrm{ov}}}{1-B_0} + t_{\mathrm{slot}},$$
$$T'_c \triangleq t_{\mathrm{coll}}, \quad \sigma \triangleq t_{\mathrm{slot}}.$$

After combining (B.1), (B.6) and (B.4), applying the equivalences just introduced and rearranging terms, one arrives at

$$a_C(0) = \bar{r} = \frac{\tau(1-p)\mathrm{E}\left[P'\right]}{(1-\tau)P_{\mathrm{empty}}\sigma + (\tau(1-p) + (1-\tau)P_{\mathrm{succ}})T'_s + ((1-\tau)P_{\mathrm{coll}} + p\tau)T'_c}.$$

Finally, employing (18) and (25)[5] to replace P_{empty}, P_{succ} and P_{coll} establishes that the mean service rate is equal to the station's saturation throughput, as given in [4], i.e., $a_C(0) = \bar{r} = S/n$, where

$$S \triangleq \frac{P_s P_{tr} \mathrm{E}\left[P'\right]}{(1-P_{tr})\sigma + P_{tr}P_s T'_s + P_{tr}(1-P_s)T'_c},$$

denotes the WLAN throughput, $P_{tr} = 1-(1-\tau)^n$ is the probability with which transmission is attempted in a considered time slot and $P_s = n\tau(1-\tau)^{n-1}/P_{tr}$ is the probability with which this attempted transmission is successful. These last two probabilities characterize events evidenced by an external system observer (monitoring n independent stations), unlike the probabilities in (25), which relate to observations by a station backing-off (monitoring $n-1$ stations).

References

[1] F. Cali, M. Conti, E. Gregory, Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit, IEEE/ACM Trans. Netw. 8 (6) (2000) 785–799.

[2] G. Bianchi, Performance Analysis of the IEEE 802.11 Distributed Coordination Function, IEEE JSAC 18 (3) (2000) 535–547.

[3] B. Li, R. Battiti, Performance Analysis of An Enhanced IEEE 802.11 Distributed Coordination Function Supporting Service Differentiation, in: Proc. QofIS, Vol. 2811, Springer LNCS, 2003, pp. 152–161.

[4] G. Bianchi, I. Tinnirello, Remarks on IEEE 802.11 DCF Performance Analysis, IEEE Commun. Lett. 9 (8) (2005) 765–767.

[5] P. Chatzimisios, A. Boucouvalas, V. Vitsas, Packet Delay Analysis of the IEEE 802.11 MAC Protocol, IEE Electronics Letters 18 (39) (2003) 1358–1359.

[6] E. Ziouva, T. Antonakopoulos, CSMA/CA performance under high traffic conditions: throughput and delay analysis, Comput. Commun. 25 (3) (2002) 313–321.

[7] A. Banchs, P. Serrano, A. Azcorra, End-to-end delay analysis and admission control in 802.11 DCF WLANs, Comput. Commun. 29 (7) (2006) 842–854.

[8] P. Raptis, A. Banchs, K. Paparrizos, A Simple and Effective Delay Distribution Analysis for IEEE 802.11, in: Proc. IEEE PIMRC, 2006.

[9] O. Tickoo, B. Sikdkar, Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC based Wireless Networks, in: Proc. IEEE INFOCOM, 2004, pp. 1404–1413.

[10] H. Zhai, Y. Kwon, Y. Fang, Performance analysis of IEEE 802.11 MAC protocols in wireless LANs, Wireless Commun. and Mobile Computing 4 (8) (2004) 917–931.

[11] H. Vu, T. Sakurai, Accurate Delay Distribution for IEEE 802.11 DCF, IEEE Commun. Lett. 10 (4) (2006) 317–319.

[12] ITU-T G.1010, End-user multimedia QoS categories (2001).

[13] L. Lin, H. Fu, W. Jia, An Efficient Admission Control for IEEE 802.11 Networks based on Throughput Analysis of Unsaturated Traffic, in: Proc. IEEE GLOBECOM, 2005, pp. 3017–3021.

[5] If a non-saturated WLAN was considered, the 'heterogeneous' analogs of these equations, as discussed in Section 5, would be employed.

[14] A. Abdrabou, W. Zhuang, Stochastic Delay Guarantees and Statistical Call Admission Control for IEEE 802.11 Single-Hop Ad Hoc Networks, IEEE Trans. Wireless Commun. 7 (10) (2008) 3972–3981.

[15] F. Kelly, Notes on effective bandwidths, in: F. Kelly, S. Zachary, I. Zeidins (Eds.), Stochastic Networks: Theory and Applications, Vol. 4, Oxford University Press, 1996, pp. 141–168.

[16] C. Chang, Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks, IEEE Trans. Autom. Control 39 (5) (1994) 913–931.

[17] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, Telecommun. Syst. 2 (1) (1993) 71–107.

[18] C. Chang, J. Thomas, Effective Bandwidth in High-Speed Digital Networks, IEEE JSAC 13 (6) (1995) 1091–1100.

[19] D. Wu, R. Negi, Effective Capacity: A Wireless Link Model for Support of Quality of Service, IEEE Trans. Wireless Commun. 2 (4) (2003) 630–643.

[20] X. Zhang, J. Tang, H. Chen, S. Ci, M. Guizani, Cross-Layer-Based Modeling for Quality of Service Guarantees in Mobile Wireless Networks, IEEE Commun. Mag. 44 (1) (2006) 100–106.

[21] G. Vecianna, G. Kesidis, J. Walrand, Resource Management in Wide-Area ATM Networks Using Effective Bandwidths, IEEE JSAC 13 (6) (1995) 1081–1090.

[22] K. Kontovasilis, N. Mitrou, Effective bandwidths for a class of non markovian fluid sources, Proc. ACM SIGCOMM, Comput. Commun. Rev. 27 (4) (1997) 263–274.

[23] A. Berman, R. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Academic Press, New York, 1979.

[24] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Nov. 1997. P802.11.

[25] IEEE 802.11b-1999 Supplement to 802.11-1999 Wireless LAN MAC and PHY specifications: Higher speed Physical Layer (PHY) extension in 2.4 GHz Band, Sept. 1999.

[26] IEEE 802.11a-1999 (8802-11:1999/Amd 1:2000(E))-specific requirements. Part 11: wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: High-speed Physical Layer in 5 GHz Band, June 2003.

[27] IEEE 802.11g-2003-Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications-Amendment 4: Further Higher-Speed Physical Layer Extension in the 2.4 GHz Band, June 2003.

[28] IEEE 802.11e-2005-Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements, Sept. 2005.

[29] Y. Peng, H. Hu, S. Cheng, K. Long, A New Self-Adapt DCF Algorithm, in: Proc. IEEE GLOBECOM, 2002, pp. 87–91.

[30] Z. Hass, J. Deng, On Optimizing the Backoff Interval for Random Access Schemes, IEEE Trans. Commun. 51 (12) (2003) 2081–2090.

[31] C. Wang, B. Li, L. Li, A New Collision Resolution Mechanism to Enhance the Performance of IEEE 802.11 DCF, IEEE Trans. on Veh. Technol. 53 (4) (2004) 1235–1246.

[32] G. Kesidis, J. Walrand, C.-S. Chang, Effective Bandwidths for Multiclass Fluids and other ATM Sources, IEEE/ACM Trans. on Networking 1 (4) (1993) 424–428.

[33] D. Malone, K. Duffy, D. Leith, Modelling the 802.11 Distributed Coordination Function in Nonsaturated Heterogeneous Conditions, IEEE/ACM Trans. Netw. 15 (1) (2007) 159–172.

[34] G. Cantieni, Q. Ni, C. Barakat, T. Turletti, Performance Analysis under Finite Load and Improvements for Multirate 802.11, Comput. Commun. 28 (10) (2005) 1095–1109.

[35] The ns-2 network simulator, www.isi.edu/nsnam/ns (1998).