Effective Capacity-based Stochastic Delay Guarantees for Systems with Time-Varying Servers, with an Application to IEEE 802.11 WLANs☆

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

[a] *National Center of Scientific Research "Demokritos", Inst. of Informatics & Telecommunications, P.O. Box 60228, GR-15310 Ag. Paraskevi, Greece.*

[b] *National & Kapodistrian University of Athens, Informatics & Telecommunications Dept., Panepistimiopolis, Illisia, Athens 15784, Greece.*

Abstract

Many network applications rely on stochastic QoS guarantees. With respect to loss-related performance, the Effective Bandwidth/Capacity theory has proved useful for calculating loss probabilities in queues with complex input- and server-processes and for formulating simple admission control tests to ensure associated QoS guarantees. This success has motivated the application of the theory for delay-related QoS too. However, up to now this application has been justified only heuristically for queues with variable service rate. The paper fills this gap by establishing rigorously that the Effective Bandwidth/Capacity theory may be used for the asymptotically correct calculation and enforcement of delay tail-probabilities in systems with vari-

☆The first and second authors' work was funded in part by the European Union through the HURRICANE project, under contract ICT-1-1.1-216006. 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)

able rate servers too. Subsequently, the paper applies the general results to IEEE802.11 WLANs, by representing each IEEE802.11 station as an On/Off server and employing the Effective Capacity function for this model. Comparison of analytical results with simulation validates the effectiveness of the On/Off IEEE802.11 model for delay-related QoS, complementing earlier results on loss-related performance.

Keywords: Effective Bandwidth, Effective Capacity, delay tail-probabilities, IEEE 802.11, QoS, Admission control, large deviations

1. Introduction

Stochastic Quality of Service (QoS) guarantees are an important ingredient of many network services. Here we focus on QoS of the form $\Pr\{D > d\} \leq e^{-\epsilon}$, where D denotes the delay experienced by traffic arriving at a queue, d is the delay threshold and ϵ represents the QoS requirement. Early relevant analyses include the development of exponential bounds of this form for GI/GI/1 FCFS queues, obtained via martingale theory [1, 2], and subsequent generalizations for Markovian arrival processes [3].

When the tail-related QoS requirement is stringent (i.e., ϵ and d are large), large deviations theory is a natural choice for obtaining the relevant results. This path has been followed for QoS requirements related to buffer overflows, leading to the, now mature, so called *Effective Bandwidth/Capacity theory*, which provides a linkage between traffic characteristics (captured by the Eff. Bandwidth function), system resources (server capacity and buffer size) and buffer content tail-probabilities. The theory was developed by many contributions over the years (see [4] for a survey in the field). It originally considered queueing systems with a constant service rate and was subsequently generalized to also address systems with time-varying servers, by introducing the Effective Capacity function to represent the server's characteristics (see, e.g., [5, 6]), analogously to the way the Effective Bandwidth function represents the input traffic.

The conceptual simplicity of the Eff. Bandwidth/Capacity theory makes it an attractive choice for coping with delay-related QoS as well. For FCFS queueing systems with a constant service rate this is directly possible, because delay probabilities of the form $\Pr\{D > d\}$ are equal to the queue length probabilities $\Pr\{Q > cd\}$, where Q and c are the queue length and the constant service rate, respectively. However, this simple equivalence does not hold when the service rate is time-varying.

Due to the prevalence of wireless networking, systems with time-varying servers are becoming all the more important. Indeed, a wireless station can be regarded as a time-varying data server, due to rate fluctuations at the Physical [7–10] or at the Medium Access Control (MAC) [11, 12] layer. Accordingly, [7] employed the Eff. Capacity function to capture the effect of a Rayleigh-fading channel on delay-related performance. However, the results developed therein cover only the restricted setting of queueing systems with constant bit rate traffic and variable service rate. Publications [8–11], as well as others, take the methodology of [7] as if it was applicable in a general setting, although no formal justification for this exists. Undoubtedly, there is a need to formalize the Eff. Bandwidth/Capacity theory for addressing delay-related QoS in the general setting of both variable input and service rates.

Fortunately, there exist suitable prior results [13–17] (not all of them directly connected with the Eff. Bandwidth/Capacity theory) that can be used for this task. By suitable extension and combination of these results, and by using an appropriate representation of the delay as the supremum of a stochastic process, this paper establishes formally that the Eff. Bandwidth/ Capacity theory may be applied for the asymptotically correct calculation of delay tail-probabilities. In particular, the paper establishes rigorously the, formerly heuristic, association of the asymptotic exponential decay rate of the delay tail-probabilities with its counterpart for the queue content tail-probabilities, through the server's Eff. Capacity function. The theory applies to queueing systems operating in either of the discrete-time or the continuous-time domain and featuring arbitrary traffic and service processes, provided these processes are independent and possess well-defined Eff. Bandwidth and Eff. Capacity functions, respectively. Besides the asymptotically tight approximation to the delay distribution's tail, the theory also suggests simple traffic admission control tests for enforcing related QoS specifications.

With the general results in hand, the paper proceeds with their application to IEEE 802.11, the prevalent standard for Wireless LANs (WLANs). To the best of the authors' knowledge, few works have been directed towards calculating delay tail-probabilities in IEEE 802.11 WLANs. Such works usually rely on classical queueing theory, thus, besides being restricted to a particular form of input traffic, they address mainly the first few moments of the delay distribution rather than tail-percentiles. The mean value of the access delay to the shared wireless medium (i.e., the first component of the overall end-to-end delay from packet generation until its single-hop delivery considered here) was calculated in [18–20], while [21–26] investigated higher-order statistics as well. References [23–25] primarily target the calculation of end-to-end related metrics, taking into account the packet waiting-time in the buffer of the IEEE 802.11 station. More specifically, [23, 24] initially 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 [23] employs an infinite G/G/1 queueing system towards calculating the mean waiting-time, while [24] relies on a simpler M/G/1/K model.

None of the results reviewed up to now are directly applicable to bursty, correlated input traffic or to QoS expressed in terms of a low probability percentile. Although the analysis in [25] is more suitable for this context, it makes the rather gross assumption that the IEEE 802.11 access delay follows a gaussian distribution when every station always has a packet to send (i.e., in saturation conditions). Furthermore, the results of [25] refer only to a restricted setting, where the data flows in the IEEE 802.11 WLAN evolve according to only two traffic profiles and one of these two types of traffic is assumed to not require any degree of QoS.

With respect to the use of the Eff. Bandwidth/Capacity theory in the context of IEEE 802.11 WLANs, publications [11, 12] are relevant. Ref. [11] models the service capacity of each IEEE 802.11 station in the WLAN as a Markov-Modulated Poisson Process (MMPP). The model is based on the assumption that the IEEE 802.11 Distributed Coordination Function (DCF) exhibits memoryless behavior 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). The establishment of the MMPP model involves a series of further approximations, towards representing the MAC dynamics in simplified terms. The Eff. Capacity of IEEE 802.11 DCF is derived from the resulting MMPP model, and is then used for the provision of stochastic delay guarantees (by employing the heuristic results of [7] and further heuristic approximations). Moreover, the traffic-control policies of [11] inherently assume that each station has a complete knowledge of the traffic load of all competing stations.

The focus of [12] is different: it targets the accurate calculation of the Eff. Capacity of each IEEE 802.11 station with the use of locally available information only, without requiring global knowledge of WLAN traffic details. Each IEEE 802.11 station is modeled as an On/Off server and the Eff. Capacity is subsequently derived from the On/Off model. 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 that leads to accurate results when the network is highly loaded. The saturation assumption is relaxed when each station measures a few model parameters (probabilities of simple local events) in a distributed manner, instead of calculating their values on the basis of the saturation assumption. This adaptation makes the model accurate for all network loads. Note that the measurement-assisted variant of the model, in common with the original saturation-based model, avoids the requirement for any knowledge about input traffic details of competing stations. This is possible because the traffic loading conditions at these stations are assessed indirectly through the measurement of the previously mentioned probabilities of local events.

This paper employs the IEEE 802.11 Eff. Capacity function proposed by [12] and uses it, in conjunction with the general results in the first part of the paper, for calculating delay tail-probabilities in IEEE 802.11 WLANs and for performing admission control to ensure a desired level of delay-related QoS. Comparison of the analytical results with simulation validates the effectiveness of the On/Off IEEE 802.11 model in the delay-related QoS context, complementing the results of [12] on loss-related performance. It is mentioned that the IEEE 802.11 Eff. Capacity model 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.

The rest of the paper is organized as follows: Section 2 discusses large deviations results for the supremum of a stochastic process. By virtue of Lindley's equation, these results lead directly to the 'ordinary' application of the Eff. Bandwidth/Capacity theory, i.e., in connection with queue content tail-probabilities. Section 2 is not just a review of preexisting results; it develops appropriately strengthened versions of these results, so that they become usable in the context of delay distributions. Section 3 firstly establishes that the delay experienced by traffic arriving at a FCFS queue has the same distribution as the supremum of a stochastic process and then applies the results of Section 2 in conjunction with preexisting results about inverse processes and process compositions, ultimately providing a rigorous justification for the use of the Eff. Bandwidth/Capacity theory in connection with delay tail-probabilities. Section 4 applies the general results to IEEE 802.11 WLANs: Firstly, the section briefly reviews the modeling of each IEEE 802.11 competing station as an On/Off server with On- and Off-sojourn times of known distributions. Subsequently, it discusses computational and algorithmic issues related to the application of the general theory of Section 3 with the particular Eff. Capacity function of this On/Off model. Section 5 validates the IEEE 802.11 model in the delay context, through comparison of analytical results with simulations. Finally, the article is concluded in Section 6.

2. Logarithmic tail-probability asymptotics for the supremum of a stochastic process

Consider a stochastic process $Y(t)$, $t \in \mathbb{T}$. The time-domain may be either discrete ($\mathbb{T} = \mathbb{Z}_+^o$) or continuous ($\mathbb{T} = \mathbb{R}_+^o$). We will be interested in asymptotics for the tail-probabilities of

$$Q \triangleq \sup_{t \in \mathbb{T}} Y(t). \tag{1}$$

In a typical application,

$$Y(t) = V(t) - C(t), \tag{2}$$

where $V(t)$ is the amount of data fed to a queue in the interval $(-t, 0]$ and $C(t)$ is the amount of data that can be processed in the same interval. Then, by Lindley's equation, Q is the queue length at time zero, provided the queueing system started operation empty an infinite amount of time ago. In

Section 3 we will encounter a stochastic process such that the supremum in (1) has the same distribution as the delay experienced by traffic arriving in a FCFS queue.

We employ throughout the following assumption about the cumulant generator of $Y(t)$, asymptotically as $t \to +\infty$:

Assumption 1.

1. *The limit*
$$u_Y(\theta) \triangleq \lim_{t\to\infty} t^{-1} \log \mathrm{E}\left[e^{\theta Y(t)}\right] \tag{3}$$
exists in the extended sense for all $\theta \in \mathbb{R}$. Let $D_Y \triangleq \{\theta : u_Y(\theta) < +\infty\}$ be the effective domain of $u_Y(\cdot)$ and denote its interior by D_Y^o.
2. *D_Y^o is nonempty and contains zero.*
3. *$u_Y(\cdot)$ is essentially smooth, namely differentiable throughout D_Y^o and steep (i.e., featuring $\lim_{n\to\infty} |u'_Y(\theta_n)| = \infty$ for any sequence θ_n taking values in D_Y^o and converging to a point on the boundary of D_Y^o).*

Since the convexity of the cumulant generator $\log \mathrm{E}\left[e^{\theta Y(t)}\right]$ is preserved by the limiting operation, $u_Y(\cdot)$ in (3) is automatically convex with $u_Y(0) = 0$. Items 1 and 2 in Assumption 1 guarantee (see, e.g., Lemma 2.3.9 in [27]) that $u_Y(\theta) > -\infty$ everywhere, so the effective domain D_Y is exactly the set where $u_Y(\cdot)$ is finite. Furthermore, by virtue of convexity D_Y is always an interval, i.e., there exist $\theta_Y^\ell < 0 < \theta_Y^u$ (because $0 \in D_Y^o$) such that $D_Y^o = (\theta_Y^\ell, \theta_Y^u)$. Each of the endpoints may be finite or infinite. If an endpoint is finite it belongs to the boundary of D_Y^o (and the steepness property in Item 3 of Assumption 1 applies to it), but it may or may not belong to D_Y. The convexity additionally implies that $u_Y(\cdot)$ is continuous in D_Y^o and upper semicontinuous at θ_Y^u and θ_Y^ℓ (i.e., $\limsup_{\theta\uparrow\theta_Y^u} u_Y(\theta) \leq u_Y(\theta_Y^u)$ and similarly for $\theta \downarrow \theta_Y^\ell$).

Typically, Assumption 1 is accompanied by the additional requirement that $u_Y(\cdot)$ is lower semicontinuous[1]. Then, whenever the upper endpoint θ_Y^u (lower endpoint θ_Y^ℓ) is finite, $u_Y(\cdot)$ is left- (right-) continuous at it. Assumption 1 together with the additional lower semicontinuity condition ensure the applicability of the Gärtner-Ellis Theorem for $Y(t)/t$ (see, e.g., Item c of Theorem 2.3.6 in [27]). However, in this paper we *do not* require $u_Y(\cdot)$ to be lower semicontinuous, the primary reason being that the stochastic process associated with the delay in FCFS queues does not always satisfy this requirement, even in cases where both the input traffic and server processes do.

With the semicontinuity assumption removed, the lower bound of the Gärtner-Ellis Theorem does not hold in its usual form anymore. Instead, a weaker form applies (see, e.g., Item b of Theorem 2.3.6 in [27]), which makes use of the exposed points of $u_Y^*(\cdot)$, the Fenchel-Legendre transform of $u_Y(\cdot)$. It will be shown that this weaker form suffices for establishing the results of interest.

We note that Assumption 1 has another implication: By Jensen's inequality $t^{-1}\log \mathrm{E}\left[e^{\theta Y(t)}\right] \geq \theta \mathrm{E}\left[Y(t)\right]/t$, so for any $\theta > 0$ one has $u_Y(\theta)/\theta \geq \limsup_{t\to\infty}(\mathrm{E}\left[Y(t)\right]/t)$. Since $u_Y'(0)$ exists, taking the limit $\theta \downarrow 0$ yields $u_Y'(0) \geq \limsup_{t\to\infty}(\mathrm{E}\left[Y(t)\right]/t)$. This result, combined with a completely analogous argument involving $\liminf_{t\to\infty}(\mathrm{E}\left[Y(t)\right]/t)$ and $\theta < 0$, establishes

[1] By convexity this property is automatically guaranteed in D_Y^o, but not necessarily on the boundary of D_Y^o.

that $\lim_{t\to\infty}(\mathrm{E}\left[Y(t)\right]/t)$ exists and

$$u_Y'(0) = \lim_{t\to\infty} \frac{\mathrm{E}\left[Y(t)\right]}{t} \triangleq \bar{r}_Y. \tag{4}$$

When $Y(\cdot)$ has stationary (or wide-sense stationary) increments $\bar{r}_Y$ is merely the mean increment per unit time.

In view of (4), one may define the 'rate' function $a_Y(\cdot)$ associated with $u_Y(\cdot)$ as follows:

$$a_Y(\theta) \triangleq \begin{cases} u_Y(\theta)/\theta, & \theta \in D_Y - \{0\}, \\ \bar{r}_Y, & \theta = 0. \end{cases} \tag{5}$$

Since $u_Y(\cdot)$ is convex with $u_Y(0) = 0$, it may be shown (see, e.g., Lemma 2.1 in [28]) that $a_Y(\cdot)$ is increasing in D_Y and continuous in D_Y^o.

Now consider

$$\theta_Y^* \triangleq \sup\{\theta : u_Y(\theta) \leq 0\}. \tag{6}$$

Since $u_Y(0) = 0$, one always has $\theta_Y^* \geq 0$. The following lemma summarizes relevant facts:

Lemma 1.

1. *$u_Y(\theta) > 0$ for all $\theta > \theta_Y^*$ (this holds with $u_Y(\theta) = +\infty$ if $\theta \notin D_Y$).*
2. *If $\theta_Y^* > 0$, then $u_Y(\theta) \leq 0$ for all $0 < \theta < \theta_Y^*$.*
3. *If $\theta_Y^* \in D_Y^o$ then:*
   (a) *θ_Y^* is a root of $u_Y(\cdot)$, i.e., $u_Y(\theta_Y^*) = 0$, and $u_Y'(\theta) > 0$ for all $\theta \in D_Y^o$ such that $\theta > \theta_Y^*$.*
   (b) *Furthermore, if there exists $\theta_o > 0$ such that $u_Y(\theta_o) < 0$, then θ_Y^* is the unique positive root of $u_Y(\cdot)$ and $u_Y'(\theta_Y^*) > 0$.*

4. *If $\bar{r}_Y < 0$, there exists $\theta_o > 0$ such that $u_Y(\theta_o) < 0$, so $\theta_Y^* > 0$. Thus, $\theta_Y^* = 0$ implies $\bar{r}_Y \geq 0$.*

Lemma 1 is a consequence of the convexity of $u_Y(\cdot)$ and Assumption 1. The complete proof is in Appendix A.

The quantity θ_Y^* is intimately connected with the tail-probabilities of Q in (1). Indeed, under appropriate conditions, $\lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\} = -\theta_Y^*$. Important results related to this asymptotic expression appear in [13–15]. The result in [13] treats discrete-time processes ($\mathbb{T} = \mathbb{Z}_+^o$) and requires a set of assumptions more restrictive than Assumption 1 (namely, that $D_Y = \mathbb{R}$ and that $\mathrm{E}\left[e^{\theta Y(t)}\right]$ satisfies additional boundedness conditions for all $t \in \mathbb{T}$). Ref. [14] is more general, but still addresses mostly the discrete-time case. Also, it requires that $\theta_Y^* \in D_Y^o$ (treating essentially the case covered by Item 3.b of Lemma 1). The result in [15] is the most general: It considers asymptotic cumulant generators of the form $\lim_{t\to\infty} v_t^{-1} \log \mathrm{E}\left[e^{\theta v_t Y(t)/a_t}\right]$, which generalize the linear scaling $v_t = a_t = t$ addressed in [13, 14] and used here, and also provides results for continuous-time processes ($\mathbb{T} = \mathbb{R}_+^o$) through additional local regularity assumptions for these processes. However, the results in [15] are expressed as separate upper and lower bounds for the tail-probabilities and these bounds are not shown to be always equal. Moreover, all mentioned results of [13–15] require that $u_Y(\cdot)$ be lower semi-continuous (the result in [13] implicitly so, by demanding $D_Y = \mathbb{R}$). To address these restrictions, we now provide the following slight strengthening of Theorems 2.1 and 2.2 in [15] for the special case $v_t = a_t = t$.

Theorem 1. *Let Assumption 1 hold. For Item 2 of the theorem only, if $\mathbb{T} = \mathbb{R}_+^o$ additionally assume that, either Hypothesis 2.3 in [15] holds, or that*

$Y(\cdot)$ is the difference of two independent processes, each having nonnegative increments and possessing an asymptotic cumulant generator as in (3). Let θ_Y^* be as in (6). Then, with Q as in (1),

1. $\liminf_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \geq -\theta_Y^*$. *Thus, if* $\theta_Y^* = 0$ *then* $\lim_{b\to\infty} \log \Pr\{Q > b\} = 0$.
2. *If there exists $\theta_o > 0$ such that $u_Y(\theta_o) < 0$ (in which case necessarily $\theta_Y^* > 0$) then* $\lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\} = -\theta_Y^*$.

The proof, to be found in Appendix B, makes use of Theorems 2.1 and 2.2 in [15] (the first of these modified to reflect the weaker Gärtner-Ellis lower bound) and Lemma 1. The method of proof makes clear that the requirement for steepness in Item 3 of Assumption 1 is only necessary when $\theta_Y^* \notin D_Y^o$. If $\theta_Y^* \in D_Y^o$, the results of Theorem 1 hold also when $u_Y(\cdot)$ is differentiable but non-steep.

Item 2 of Theorem 1 may be regarded as a proof that the lower and upper bounds of Theorems 2.1 and 2.2 in [15], as specialized for the linear scaling, always coincide. With respect to the additional assumptions required for continuous-time processes, it is noted that a sufficient (but not necessary) condition for satisfying Hypothesis 2.3 in [15] is that the increments of $Y(\cdot)$ are bounded. The alternative condition of Theorem 1 is automatically satisfied in real-world queues where $Y(\cdot)$ is the difference of the amount of data fed to the queue minus the amount of data that can be processed. Both of these have nonnegative increments.

Item 4 of Lemma 1 ensures that whenever $\bar{r}_Y < 0$ (i.e., whenever the system is stable), Item 2 of Theorem 1 applies, so the full limit therein exists. The next corollary suggests that when the rate function $a_Y(\cdot)$ is strictly increasing, the existence of this full limit is always guaranteed:

Corollary 1. *If the assumptions in the beginning of Theorem 1 hold and furthermore $a_Y(\cdot)$ in (5) is strictly increasing, then always* $\lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\} = -\theta_Y^*$.

Proof. If $a_Y(0) < 0$, then by continuity $a_Y(\theta_o) < 0$ for $\theta_o > 0$ close enough to zero, thus $u_Y(\theta_o) = \theta_o a_Y(\theta_o) < 0$ and Item 2 of Theorem 1 applies. If $a_Y(0) \geq 0$, by strict monotonicity $a_Y(\theta) > 0, \forall\theta > 0$, so Item 1 of Theorem 1 applies with $\theta_Y^* = 0$. □

Given a stochastic process $Y(t)$, $t \in \mathbb{T}$, Theorem 1 determines the asymptotic exponential decay rate of the tail-probabilities of Q. However, many practical queueing applications call for the 'reversed' objective of ensuring that the decay rate be bounded from below[2] by some threshold determined by the QoS requirements. Corollary 2 below links such guarantees with associated conditions expressed in terms of the asymptotic cumulant generator $u_Y(\cdot)$. As we shall see later in this section, these conditions essentially pose a limit to the amount of traffic entering the queue, thus they act as natural traffic admission control tests.

Corollary 2. *Let the assumptions in the beginning of Theorem 1 hold. Then, for any $\theta > 0$:*

1. $u_Y(\theta) < 0$ *implies* $\lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \leq -\theta$.
2. $\limsup_{b\to\infty} b^{-1} \log \Pr\{Q > b\} < -\theta$ *implies that* $u_Y(\theta) \leq 0$.

[2] Decay rates are defined as positive quantities; in Theorem 1 and Corollary 1 the decay rate is θ_Y^*, not $-\theta_Y^*$.

3. *Furthermore, if* $a_Y(\cdot)$ *is strictly increasing and* $\sup D_Y \triangleq \theta_Y^u = +\infty$, *then*

$$u_Y(\theta) \leq 0 \Leftrightarrow \lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \leq -\theta$$

and equality at one side of the equivalence implies equality at the other side too.

Proof. With respect to Item 1, the condition $u_Y(\theta) < 0$ and (6) yield $\theta \leq \theta_Y^*$, so $-\theta \geq -\theta_Y^* = \lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\}$, the equality following from Item 2 of Theorem 1. For Item 2 of the corollary, combine the condition with Item 1 of Theorem 1 to get

$$-\theta_Y^* \leq \liminf_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \leq \limsup_{b\to\infty} b^{-1} \log \Pr\{Q > b\} < -\theta.$$

Therefore, $\theta < \theta_Y^*$ and $u_Y(\theta) \leq 0$ by Item 2 of Lemma 1. With respect to Item 3, the strict monotonicity of $a_Y(\cdot)$ and Corollary 1 ensure the existence of the limit at the right hand side of the equivalence. Then, the implication from the left to the right hand side follows by the reasoning used in proving Item 1. Moreover, equality at the left hand side for some $\theta > 0$ and the strict monotonicity of $a_Y(\cdot)$ imply that $\theta = \theta_Y^*$ and equality holds at the right hand side too. In the reverse direction, assume that the right hand side of the equivalence holds; then, Corollary 1 yields $\theta \leq \theta_Y^*$. If $\theta_Y^* = +\infty$ both sides of the equivalence hold with strict inequality (the left side due to the strict monotonicity of $a_Y(\cdot)$). Otherwise, $\theta_Y^* \in D_Y^o$ (due to the assumption $\sup D_Y = +\infty$), and this fact, coupled with the strict monotonicity of $a_Y(\cdot)$ make Item 3.b of Lemma 1 applicable. Therefore, $\theta < \theta_Y^*$ (resp. $\theta = \theta_Y^*$) implies that both sides of the equivalence hold with strict inequality (resp. equality). □

Item 1 of Corollary 2 establishes the sufficiency of the condition $u_Y(\theta) < 0$ for ensuring that the asymptotic exponential decay rate of the tail-probabilities of Q is bounded from below by θ. Item 2 is a partial converse, showing that the slightly more general condition $u_Y(\theta) \leq 0$ is necessary for such a bound to exist. According to Item 3, the strict monotonicity of the rate function enables a full equivalence. The requirement for $\sup D_Y = \theta_Y^u = +\infty$ in this last item is a technical condition to exclude cases with $\theta_Y^* = \theta_Y^u < +\infty$ and $u_Y(\theta_Y^*) \neq 0$. (In such cases the equivalence could break for $\theta = \theta_Y^*$ but would still hold for $\theta < \theta_Y^*$.) As the proof reveals, the technical condition is needed only for the implications from the right to the left part of the equivalence. It is noted that the first two items of Corollary 2 may be regarded as generalizations of Theorem 3.8 and part (i) of Theorem 3.9 in [13] to a broader context.

We close this section by linking its results with the 'ordinary' application of the Eff. Bandwidth/Capacity theory for queue content tail-probabilities. As already mentioned, in this case the process $Y(t)$ has the form (2). If the traffic process $V(t)$ and server process $C(t)$ are independent and if both have asymptotic cumulant generators, $u_V(\cdot)$ and $u_C(\cdot)$, of the form (3), then

$$u_Y(\theta) = u_V(\theta) + u_C(-\theta). \tag{7}$$

Moreover, if each of the traffic and server processes satisfies Assumption 1, then their difference (2) does too. Also, for the case $\mathbb{T} = \mathbb{R}_+^o$, if both $V(t)$ and $C(t)$ either satisfy Hypothesis 2.3 of [15] or have nonnegative increments, then their difference (2) satisfies the additional assumption of Theorem 1 and the results of this section apply.

In view of (4), the asymptotic mean rate takes the intuitive form $\bar{r}_Y =$

$\bar{r}_V - \bar{r}_C$, so the 'stability condition' $\bar{r}_Y < 0$ (see Item 4 of Lemma 1 and the comments before Corollary 1) translates to the usual queue stability condition. Similarly, by virtue of (5) and (7), the rate function takes the form $a_Y(\theta) = a_V(\theta) - a_C(-\theta)$; the function $a_V(\cdot)$ is the Eff. Bandwidth function, while $a_C(\cdot)$ is the Eff. Capacity function. In connection with Corollary 1 and Item 3 of Corollary 2, it is noted that if at least one of the Eff. Bandwidth and the Eff. Capacity functions is strictly increasing, then $a_Y(\cdot)$ also is.

Given the particular form of the rate function in the queueing context, (6) suggests that, whenever $\theta_Y^* > 0$, it may be determined as $\theta_Y^* = \sup\{\theta : a_V(\theta) \leq a_C(-\theta)\}$, i.e., as the maximum parameter θ for which the Eff. Bandwidth does not exceed the Eff. Capacity. Under this point of view, $a_V(\theta)$ $\big($resp. $a_C(-\theta)\big)$ is to be interpreted as the bandwidth requirements of the traffic (resp. the server's capacity) with respect to parameter θ and then θ_Y^* emerges as the maximal parameter value that satisfies the corresponding "generalized queue stability condition". Furthermore, in the usual (but not exclusively encountered in applications) case when $\theta_Y^* \in D_Y^o$, the asymptotic decay rate satisfies $a_V(\theta_Y^*) = a_C(-\theta_Y^*)$.

Similar comments apply with respect to Corollary 2: the condition $u_Y(\theta) < 0$ (resp. $u_Y(\theta) \leq 0$) therein translates again to the abovementioned "generalized queue stability condition", viz., $a_V(\theta) < a_C(-\theta)$ $\big($resp. $a_V(\theta) \leq a_C(-\theta)\big)$. These conditions are very suitable for admission control tests. The complexity of these tests does not grow even when the traffic is a complex superposition of independent traffic streams, as in this case the overall Eff. Bandwidth function is merely the sum of the Eff. Bandwidth functions of the constituent streams. The value of θ to be used in the tests is determined as follows: The stochastic, loss-related QoS specification dictates that the queue content should not exceed some given level x (this event being taken as a proxy to overflows in a system with finite buffer of size x) with probability higher than $e^{-\epsilon}$. Provided that both x and ϵ are large maintaining a finite ratio, the QoS specification leads to $-\epsilon/x \geq x^{-1} \log \Pr\{Q > x\} \approx \lim_{b\to\infty} b^{-1} \log \Pr\{Q > b\}$, so $\theta = \epsilon/x$ should be used in the admission control tests.

3. Effective Bandwidth/Capacity theory for delay probabilities in FCFS queues

We now revisit the queueing context discussed in the last part of the previous section. We assume that the queue operates according to the FCFS policy and let D stand for the delay experienced by data entering the queue at $t = 0$ (an infinite amount of time after the system has started operation). Moreover, we employ the following:

Assumption 2.

1. *The traffic process $V(t)$ and the server process $C(t)$, $t \in \mathbb{T}$, are mutually independent.*
2. *$C(t)$, $t \in \mathbb{T}$, has nonnegative and stationary increments.*
3. *Each of $V(t)$ and $C(t)$, $t \in \mathbb{T}$, satisfy Assumption 1 with asymptotic cumulant generators $u_V(\cdot)$ and $u_C(\cdot)$, respectively.*
4. *Furthermore, $u_C(\cdot)$ is lower semicontinuous.*
5. *For continuous time processes only ($\mathbb{T} = \mathbb{R}_+^o$): $V(t)$, $t \in \mathbb{T}$, also has nonnegative increments.*

Items 1 and 3 of this assumption are the requirements used in the last part of Section 2 for applying the Eff. Bandwidth/Capacity theory to the tail-probabilities of the queue content distribution. As will be discussed later, Items 2 and 4 (also Item 5, when $\mathbb{T} = \mathbb{R}_+^o$) are additional requirements to ensure that D has the same distribution as the supremum of a stochastic process featuring a well-behaved asymptotic cumulant generator, so that the results of Section 2 may be applied.

Indeed, the nonnegativity of increments in Item 2 of the assumption ensures that $C(\cdot)$ possesses an inverse process [17], namely

$$T(v) \triangleq \inf\{s \geq 0 : C(s) \geq v\}. \tag{8}$$

The following result links the inverse process $T(\cdot)$ and the traffic process $V(\cdot)$ to the delay D, analogously to the way Lindley's equation links the workload process to the queue content:

Proposition 1. *If Items 1 and 2 in Assumption 2 hold, then $D =_d \sup_{t\in\mathbb{T}} Z(t)$, where*

$$Z(t) \triangleq T\big(V(t)\big) - t, \quad t \in \mathbb{T}. \tag{9}$$

Proof. Let $C(t_1, t_2]$ denote the amount of data that can be processed in the time-interval $(t_1, t_2]$. With this notation, $C(t) \triangleq C(-t, 0]$. We now show that $\Pr\{D > d\} = \Pr\{\sup_{t\in\mathbb{T}} Z(t) > d\}$, for all $d \in \mathbb{R}$. This is immediate for $d < 0$, since both D and $\sup_{t\in\mathbb{T}} Z(t)$ are nonnegative (the second one by construction, in view of (9) and the fact that $V(0)$ and $T(0)$ are zero w.p. 1).

For $d \geq 0$, one has:

$$\begin{aligned}
\Pr\{D > d\} &= \Pr\{C(0,d] < Q\} \\
&= \Pr\big\{C(0,d] < \sup_{t\in\mathbb{T}}\{V(t) - C(t)\}\big\} \\
&= \Pr\big\{\cup_{t\in\mathbb{T}}\{C(0,d] < V(t) - C(-t,0]\}\big\} = \Pr\big\{\cup_{t\in\mathbb{T}}\{C(-t,d] < V(t)\}\big\} \\
&= \Pr\big\{\cup_{t\in\mathbb{T}}\{C\big(-(t+d),0\big] < V(t)\}\big\} = \Pr\big\{\cup_{t\in\mathbb{T}}\{C(t+d) < V(t)\}\big\} \\
&= \Pr\big\{\cup_{t\in\mathbb{T}}\{T\big(V(t)\big) > t+d\}\big\} = \Pr\big\{\cup_{t\in\mathbb{T}}\{Z(t) > d\}\big\} = \Pr\{\sup_{t\in\mathbb{T}} Z(t) > d\}.
\end{aligned}$$

The first equality above is due to the FCFS policy, while the second follows from Lindley's equation (see (1) and (2)). The fourth equality is a result of stationarity, which implies that the joint distribution of the increments of the server process is invariant to a translation of time by $-d$. Finally, the fifth equality, which proves the result, is a direct consequence of the definition $T(\cdot)$ in (8), Item 1 of Assumption 2 and (9). □

By its definition, the inverse process $T(\cdot)$ has nonnegative and stationary increments, inheriting these properties from $C(\cdot)$. Thus, Item 5 in Assumption 2 is enough to guarantee that nonnegativity of increments is also a feature of $T(V(\cdot))$, so $Z(\cdot)$ in (9) is the difference of two independent processes, each with nonnegative increments, and the additional requirement of Theorem 1 when $\mathbb{T} = \mathbb{R}_+^o$ is satisfied.

In view of (9) we proceed to determine the asymptotic cumulant generator $u_Z(\cdot)$ and to check whether Assumption 1 is satisfied. Clearly (see (3) and (9)),

$$u_Z(\xi) = u_{T\circ V}(\xi) - \xi, \quad \forall \xi \in \mathbb{R}, \tag{10}$$

provided the asymptotic cumulant generator $u_{T\circ V}(\cdot)$ is well defined. (Here and in the following we employ the usual composition operator notation

$T \circ V(\cdot) \triangleq T\big(V(\cdot)\big)$.) Since $C(\cdot)$, thus also $T(\cdot)$, is independent from $V(\cdot)$, Theorem 5 in [16] yields

$$u_{T\circ V}(\xi) = \begin{cases} u_V\big(u_T(\xi)\big), & \xi \in D^o_{T\circ V} = D^o_V \cap u_T(D^o_T), \\ +\infty, & \mathbb{R} \setminus \overline{D^o_{T\circ V}}\,. \end{cases} \tag{11}$$

Note that, although the theorem clearly determines the values of $u_{T\circ V}(\cdot)$ in the interior of its effective domain (which is also designated), it cannot specify what happens on the boundary of $D_{T\circ V}$. Indeed, if D_V is closed at any of its endpoints, e.g., $D_V = (\theta^\ell_V, \theta^u_V]$, and if there exists $\xi_0 \in D^o_T$, such that $u_T(\xi_0) = \theta^u_V$, it may happen that $u_{T\circ V}(\xi_0) = +\infty$, although $u_V(u_T(\xi_0))$ is finite and, moreover, $\lim_{\xi\uparrow\xi_0} u_{T\circ V}(\xi)) = u_V(u_T(\xi_0)) < u_{T\circ V}(\xi_0)$. Consequently, $u_{T\circ V}(\cdot)$ (thus also $u_Z(\cdot)$) may fail to be lower semicontinuous even when both $u_V(\cdot)$ and $u_T(\cdot)$ are. This observation provided the primary motivation in this paper for developing Theorem 1, which avoids the dependence on lower semicontinuity assumptions.

We now express $u_T(\cdot)$ in terms of $u_C(\cdot)$. In preparing for this, we observe that the nonnegativity in Item 2 of Assumption 2 implies that $u_C(\cdot)$ is nondecreasing throughout the interior of its effective domain $D^o_C = (-\infty, \theta^u_C)$, where $\theta^u_C > 0$ by virtue of Item 3 in the assumption. Let $\hat{\theta}^\ell_C \triangleq \inf\{\theta : u_C(\theta) > u_C(-\infty)\}$; by convexity, $u_C(\theta)$ is strictly increasing for all $\theta > \hat{\theta}^\ell_C$ (and constant and equal to $u_C(-\infty)$ for all $\theta \le \hat{\theta}^\ell_C$). Although in most applications $u_C(\cdot)$ is strictly increasing throughout D^o_C and $\hat{\theta}^\ell_C = -\infty$, in the interest of generality we also consider the possibility that $\hat{\theta}^\ell_C$ is finite. The degenerate case of a null service process (i.e., $C(t) = 0$ w.p. 1, for all $t \in \mathbb{T}$) is excluded from further consideration, because in this case $T(v) = +\infty$ w.p. 1 for all $v > 0$, so $D = +\infty$ w.p. 1 as well, (unless the traffic process is null too). With the degenerate case excluded, one always has $\hat{\theta}_C^\ell < 0$ (because $u_C'(0) = \bar{r}_C > 0$).

In the context just discussed one may employ Theorem 1 in [17] (whose application requires Items 3 and 4 of Assumption 2) to obtain $u_T(\cdot)$ as follows:

$$u_T(\xi) = \begin{cases} -\theta_C^u, & \xi \leq -u_C(\theta_C^u), \\ -u_C^{-1}(-\xi), & -u_C(\theta_C^u) < \xi < -u_C(\hat{\theta}_C^\ell), \\ +\infty, & \xi > -u_C(\hat{\theta}_C^\ell) = -u_C(-\infty). \end{cases} \tag{12}$$

The theorem ensures that $u_T(\cdot)$ satisfies Assumption 1, inheriting this property from $u_C(\cdot)$. It is noted that, whenever $\hat{\theta}_C^\ell > -\infty$, the value of $u_T\big(-u_C(\hat{\theta}_C^\ell)\big)$ is ambiguous (it may be equal to $+\infty$, or to $-\hat{\theta}_C^\ell$). However, in all cases $D_T^o = \big(-\infty, -u_C(\hat{\theta}_C^\ell)\big)$.

It is now straightforward to combine (12) with (11) for determining $D_{T\circ V}^o$ (which is equal to D_Z^o, due to (10)). Indeed, with $D_V^o = (\theta_V^\ell, \theta_V^u)$ and $D_{T\circ V}^o = D_Z^o \triangleq (\xi_Z^\ell, \xi_Z^u)$, one has

$$\xi_Z^\ell = \begin{cases} -\infty, & \theta_V^\ell \leq -\theta_C^u, \\ -u_C(-\theta_V^\ell), & \theta_V^\ell > -\theta_C^u, \end{cases} \qquad \text{and} \qquad \xi_Z^u = -u_C\big(-\min\{-\hat{\theta}_C^\ell,\, \theta_V^u\}\big). \tag{13}$$

When $V(\cdot)$ has nonnegative increments, the relation for ξ_Z^ℓ above reduces always to the first branch, because $\theta_V^\ell = -\infty$. This is consistent with the earlier observation that Item 5 of Assumption 2 is sufficient for ensuring that $T \circ V(\cdot)$ also has nonnegative increments.

Up to this point we have established that $u_Z(\cdot)$ is well defined (through (10), (11) and (12)) and that the interior of its effective domain is nonempty and contains zero. The differentiability of $u_Z(\cdot)$ follows from the differentiability of $u_V(\cdot)$ and $u_C(\cdot)$, itself assured by Item 3 of Assumption 2. Moreover, Appendix C shows that $u_Z(\cdot)$ is steep. Therefore, $u_Z(\cdot)$ fulfills all the requirements for the validity of Assumption 1 and it becomes possible to apply Theorem 1 for the tail-probabilities of the delay D. According to the theorem, the asymptotic decay rate of these probabilities is

$$\xi_Z^* \triangleq \sup\{\xi : u_Z(\xi) \leq 0\}. \tag{14}$$

As discussed in Section 2, always $\xi_Z^* \geq 0$. Furthermore, (10), (11) and (12) suggest that, when $\xi \geq 0$ (actually also for negative values in a range), the form of $u_Z(\xi)$ simplifies to

$$u_Z(\xi) = \begin{cases} u_V\big(-u_C^{-1}(-\xi)\big) - \xi, & \max\{\xi_Z^\ell, -u_C(\theta_C^u)\} < \xi < \xi_Z^u, \\ +\infty, & \xi > \xi_Z^u, \end{cases} \tag{15}$$

with ξ_Z^ℓ and ξ_Z^u as in (13). However, it is not even necessary to apply (15) and (14) for determining ξ_Z^*, because the following result shows how to obtain it from the corresponding decay rate of the queue content tail-probabilities:

Theorem 2. *If Assumption 2 holds, then $\xi_Z^* = -u_C(-\theta_Y^*)$, with ξ_Z^* as in* (14) *and θ_Y^* as in* (6).

Proof. For all $0 \leq \xi < \xi_Z^* \leq -u_C(\hat{\theta}_C^\ell)$ (see (13) and (12)), the function $-u_C^{-1}(-\cdot)$ is strictly increasing. Similarly, the inverse function $-u_C(-\cdot)$ is also nondecreasing and continuous. Thus, using the one-to-one transformation $\theta = -u_C^{-1}(-\xi)$ together with (13), (14) and (15), we obtain $\xi_Z^* = \sup_{\theta\in\Theta}\{-u_C(-\theta)\} = -u_C(-\sup\Theta)$, where, using also (7), $\Theta \triangleq \big\{\theta : \theta < \min\{-\hat{\theta}_C^\ell, \theta_V^u\}, u_Y(\theta) = u_V(\theta)+u_C(-\theta) \leq 0\big\}$. Therefore, $\sup\Theta = \min\{-\hat{\theta}_C^\ell, \theta_V^u, \theta_Y^*\} = \min\{-\hat{\theta}_C^\ell, \theta_Y^*\}$, because always $\theta_Y^* \leq \theta_V^u$.

If $\theta_Y^* \leq -\hat{\theta}_C^\ell$ there is nothing further to prove. In the complementary case, $-\theta_Y^* < \hat{\theta}_C^\ell$ and since $u_C(\cdot)$ is constant for all $\theta \leq \hat{\theta}_C^\ell$, it follows that $\xi_Z^* = -u_C(\hat{\theta}_C^\ell) = -u_C(-\theta_Y^*)$, yielding the same result. □

In a sense, Theorem 2 provides a natural generalization over systems with a constant server rate c. In such systems always $\Pr\{D > d\} = \Pr\{Q > cd\}$, so the asymptotic decay rates are necessarily linked by the relation $\xi_Z^* = c\theta_Y^*$. The theorem reflects this because, when the service rate is constant $u_C(\theta) = c\theta$. Moreover, in a general setting with variable service rate, $-u_C(-\theta_Y^*) = a_C(-\theta_Y^*)\theta_Y^*$, so the system "appears" as if it featured a constant server rate equal to $a_C(-\theta_Y^*)$. This is consistent with the discussion at the end of Section 2 about the role of the Eff. Bandwidth value $a_V(\theta_Y^*)$ and the Eff. Capacity value $a_C(-\theta_Y^*)$ as descriptors of the traffic's bandwidth requirements and the server's capacity, respectively.

By Item 3 of Lemma 1 and (7), $u_V(\theta_Y^*) + u_C(-\theta_Y^*) = 0$ whenever $\theta_Y^* \in D_Y^o$, so ξ_Z^* may be expressed in this case through the asymptotic cumulant generator of the traffic process as $\xi_Z^* = u_V(\theta_Y^*)$. However, the result of Theorem 2 and its interpretation discussed in the previous paragraph hold even in settings where $\theta_Y^* \notin D_Y^o$ with $u_Y(\theta_Y^*) \neq 0$. Such cases are not necessarily exotic; for an example see Appendix D.

It is noted that, besides Theorem 2, the intimate relationship between queue content and delay is manifested in other aspects too. Indeed, (15), (7) and (4) suggest that $\bar{r}_Z = \bar{r}_V/\bar{r}_C - 1$, so $\bar{r}_Z < 0$ iff $\bar{r}_Y = \bar{r}_V - \bar{r}_C < 0$, i.e., the system is stable in terms of the queue content if and only it is stable in terms of the delay. Similarly, (15), (5) and the strict monotonicity of $-u_C^{-1}(-\cdot)$ reveal that if at least one of the Eff. Bandwidth function $a_V(\cdot)$

and the Eff. Capacity function $a_C(\cdot)$ is strictly increasing, then both the rate functions $a_Y(\cdot)$ and $a_Z(\cdot)$ will also be strictly increasing and, by Corollary 1, the tail-probabilities of the queue content and of the delay will both possess a full logarithmic limit.

As with the queue content, we now consider admission control for ensuring delay-related QoS guarantees. For this purpose we can apply Corollary 2 to the process $Z(t)$, $t \in \mathbb{T}$, and the quantities associated with it. Then, in order to ensure that the decay rate of the delay tail-probabilities is bounded below by some $\xi > 0$, the admission control condition $u_Z(\xi) < 0$ (or $u_Z(\xi) \leq 0$, if Item 3 in the corollary applies) must be tested. In light of (15), this is equivalent to setting

$$\theta(\xi) \triangleq -u_C^{-1}(-\xi) \tag{16}$$

and then testing for $u_V(\theta(\xi)) < \xi$. The test may also be expressed in terms of the Eff. Bandwidth function as

$$a_V(\theta(\xi)) < \xi/\theta(\xi). \tag{17}$$

Obviously, these tests are no different than $u_Y(\theta(\xi)) = u_V(\theta(\xi)) + u_C(-\theta(\xi)) < 0$ for the first form and $a_V(\theta(\xi)) < a_C(-\theta(\xi))$ for the second. These alternate forms (together with the fact $\xi = -u_C(-\theta(\xi))$ and Theorem 2) emphasize the connection with the queue length context, but are computationally less appealing than their previous counterparts.

The value of the parameter ξ to employ in the tests is determined in a way analogous to the one used for loss-related QoS requirements. This time the QoS specification dictates that the delay should not exceed some given threshold τ with probability higher than $e^{-\epsilon}$. Provided that both τ

and ϵ are large maintaining a finite ratio, the QoS specification leads to $-\epsilon/\tau \geq \tau^{-1} \log \Pr\{D > \tau\} \approx \lim_{d\to\infty} d^{-1} \log \Pr\{D > d\}$, so $\xi = \epsilon/\tau$ should be used in the admission control tests.

There is one further thing that requires attention: Theorem 2 suggests that the asymptotic exponential decay rate of the delay tail-probabilities cannot exceed $-u_C(-\infty)$. Thus, if this quantity is finite, any QoS specification greater than it cannot be satisfied, regardless of how low the input traffic may be[3]. In light of these comments, the admission control tests presented before should be preceded by the test $\xi < -u_C(-\infty)$. If this test fails, then the admission control test fails too, otherwise the normal test described before is applied. Note that the extra test just discussed is never required in settings where the server rate is always maintained greater than a positive threshold, because in this case it is guaranteed that $u_C(-\infty) = -\infty$ and any degree of QoS may be accommodated (provided the input traffic is suitably restricted). However, if the server rate may attain zero values over some period of time, a finite value of $u_C(-\infty)$ may indeed occur. We will encounter this phenomenon in the next section, where IEEE 802.11 stations are modeled as On/Off servers.

4. The Effective Capacity of IEEE 802.11 stations

We now apply the general results to IEEE 802.11 WLANs. Subsection 4.1 describes the representation of IEEE 802.11 stations in the WLAN as On/Off

[3] An alternative way of seeing this effect is through (13), which shows that when $\xi > -u_C(-\infty) = -u_C(\hat{\theta}^{\ell}_C)$ then also $\xi > \xi^u_Z$, so $u_Z(\xi) = +\infty$ and the conditions in Item 1 (or Item 3) of Corollary 2 cannot be satisfied.

servers, according to [12]. Subsection 4.2 discusses the use of the Eff. Capacity resulting from this On/Off model in connection with delay-related QoS requirements. Subsection 4.1 is limited only to the material absolutely necessary for stating the On/Off model and for associating it with the IEEE 802.11 MAC protocol (with which the reader is assumed to be familiar). For further details, proofs and additional insight, the reader is referred to [12].

4.1. Representing mobile stations as On/Off servers

Because of the Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) access algorithm used by the IEEE 802.11 protocol, a mobile station behaves as a On/Off server. The server is On (at a rate equal to the channel bit rate $\hat{r}$) when transmitting successfully the payload of a packet. In all other states of the IEEE 802.11 protocol (station backing-off, colliding with other stations, or doing overhead operations before or after a successful transmission, e.g., RTS/CTS or ACK), the server is Off. Note that the On/Off model puts in the Off period all signaling and other overhead operations (including the transmission of the packet's header), and thus assigns a zero service rate to them, even though the IEEE 802.11 station actually transmits signaling data and/or packet header (at the channel rate $\hat{r}$) during some of these operations. This arrangement is appropriate for representing the service rate available to higher layers of the protocol stack.

Let T_{on} and T_{off} stand for the On- and Off-sojourn times, respectively. The moment generator of T_{on} is simply

$$\gamma_{\mathrm{on}}(\omega) \triangleq \mathrm{E}\left[e^{\omega T_{\mathrm{on}}}\right] = \mathrm{E}\left[e^{\omega P/\hat{r}}\right], \tag{18}$$

where P is the payload size of the packet being transmitted. When packets have a constant payload, T_{on} is a deterministic random variable. The moment generator of T_{off} reads

$$\gamma_{\text{off}}(\omega) \triangleq \mathrm{E}\left[e^{\omega T_{\text{off}}}\right] = e^{\omega t_{\text{over}}}\Big(B_o + (1-B_o)e^{\omega t_{\text{slot}}}\gamma_{\text{bo}}(\omega)\Big). \tag{19}$$

This equation reflects the fact that, when the station, after a successful transmission, draws a 0^{th}-stage backoff window equal to zero (an event of probability B_o), the Off period simply lasts a constant time t_{over}, equal to the time required for performing the overhead operations before and after the successful transmission. In the complementary event, with probability $1 - B_o$, the Off period additionally includes the constant time t_{slot}, required for initially decrementing the backoff counter by one, plus the time spent by the station in backoff mode. The moment generator for this backoff time is

$$\gamma_{\text{bo}}(\omega) = \frac{g_o\big(\gamma_s(\omega)\big) - B_o}{\gamma_s(\omega)(1-B_o)} \sum_{l=0}^{\infty}\Big((1-p)p^l e^{l t_{\text{coll}}\omega} \prod_{j=1}^{l} g_j\big(\gamma_s(\omega)\big)\Big), \tag{20}$$

where t_{coll} is the constant time required for detecting a collision and $g_j(\cdot)$ is the generator function of the backoff window W_j drawn at the j^{th} backoff stage, $j \geq 0$. Always, $B_o = g_0(0)$. Eq. (20) allows the use of general backoff window distributions. For the special uniform distributions employed by the IEEE 802.11 standard, $g_j(z) = {w_j}^{-1}\sum_{l=0}^{w_j-1} z^l = {w_j}^{-1}(z^{w_j}-1)/(z-1)$, where $w_j = 2^{\min\{j,m\}}w_o, j \geq 0$ and m is the threshold of the backoff stage index beyond which the backoff window distributions remain invariant.

The quantity p in (20), called conditional collision probability, denotes the collision probability observed by a packet attempting transmission. The value of p is obtained by solving the system of equations

$$1-p=(1-\tau)^{n-1}, \qquad \tau=\Big[1+(1-p)\big(\frac{\mathrm{E}\left[W_o\right]}{1-B_o}-1+\sum_{i=1}^{\infty}p^i\mathrm{E}\left[W_i\right]\big)\Big]^{-1} \tag{21}$$

for the conditional collision probability p and the transmission probability τ [29]. In (21), n is the number of competing stations and $\mathrm{E}\left[W_i\right]=g_i'(1)$ is the mean backoff window at the i^{th} backoff stage.

Finally, the function $\gamma_s(\omega)$ appearing in (20) is the moment generator of the time required for the reduction of the backoff counter by one, viz.,

$$\gamma_s(\omega)=P_{\text{coll}}e^{\omega t_{\text{coll}}}+P_{\text{empty}}e^{\omega t_{\text{slot}}}+P_{\text{succ}}\frac{(1-B_o)\gamma_{\text{on}}(\omega)e^{\omega t_{\text{over}}}}{1-B_o\gamma_{\text{on}}(\omega)e^{\omega t_{\text{over}}}}e^{\omega t_{\text{slot}}}, \tag{22}$$

where

$$P_{\text{succ}}=(n-1)\tau(1-\tau)^{n-2}, \quad P_{\text{empty}}=(1-\tau)^{n-1}, \quad P_{\text{coll}}=1-P_{\text{succ}}-P_{\text{empty}}, \tag{23}$$

are the probabilities with which a successful transmission, an empty slot and a collision, respectively, are observed by a station backing-off (this station observing $n-1$ other independent stations).

It is noted that when the backoff stage index threshold m is finite, the infinite sum in the expression for τ within (21) specializes to $\sum_{i=1}^{m-1}p^i\mathrm{E}\left[W_i\right]+p^m\mathrm{E}\left[W_m\right]/(1-p)$. Similarly, the infinite sum in (20) may also be written in closed form. At this point it is noted that the constant times t_{slot}, t_{over} and t_{coll}, used in (19), (20) and (22), are simple functions of basic MAC parameters specified by the standard; for details, see [12].

The formulation of (21) assumes saturation conditions, in which all other competing stations always have a packet to send. This is a conservative assumption, suitable for highly loaded networks. The dependence of $\gamma_{\text{off}}(\cdot)$ on the saturation assumption is only through the conditional collision probability p, used in (20) and the probabilities P_{succ}, P_{empty} and P_{coll} employed by (22). Under non-saturation conditions these parameters retain their meaning, but take different values. Thus, if each mobile station assesses these probabilities by direct measurement, rather than computing them through (21) and (23), the model works well in all settings, lightly loaded ones included.

4.2. Using the IEEE 802.11 Effective Capacity for delay-related QoS

The moment generators of the On- and Off-periods of the IEEE 802.11 model in (18) and (19) have open effective domains, D_{on} and D_{off}, respectively, both of them containing zero. Indeed, $D_{\text{on}} = \mathbb{R}$, because the size of transmitted packets is always nonnegative and upper bounded by the maximum PDU size. Also, $D_{\text{off}} = (-\infty, \omega^*_{\text{off}})$, where ω^*_{off} is determined by the requirement that the infinite sum in (20) converges. If the threshold stage m is finite, as in the standard, then the infinite sum in (20) can be expressed in closed form and ω^*_{off} is the unique positive value of ω satisfying $g_m\big(\gamma_s(\omega)\big)e^{\omega t_{\text{coll}}} = 1/p$.

Given the properties of the effective domains just mentioned, it is possible to determine the asymptotic cumulant generator (3) of the On/Off model, by employing general results for semi-Markovian models [28]. According to these results, $u_C(\cdot)$ is a finite and analytic function in the entire set of real numbers and can be derived by means of an implicit function problem, which, for the On/Off case of interest here, takes the form

$$f(\theta, u_C(\theta)) = 0, \qquad f(\theta, u) \triangleq \log \gamma_{\text{on}}(\hat{r}\theta - u) + \log \gamma_{\text{off}}(-u). \tag{24}$$

In light of the previous comments and the fact that the rate process of the On/Off model is nonnegative and stationary, Items 2–4 of Assumption 2 are seen to be satisfied. Moreover, in connection with Corollary 1 and Item 3 in Corollary 2, it is mentioned that, since T_{off} is not constant w.p. 1, it is guaranteed [28] that the Eff. Capacity function $a_C(\cdot)$ is strictly increasing, with $a_C(0) = \bar{r}_C$ and $\lim_{\theta \to -\infty} a_C(\theta) = 0$.

Assume now that the IEEE 802.11 station is loaded by a traffic process with nonnegative increments and an asymptotic cumulant generator $u_V(\cdot)$ satisfying Assumption 1. Then Assumption 2 is satisfied in full and the results of Section 3 can be used. In view of Theorem 2, the asymptotic decay rate ξ_Z^* of the delay tail-probabilties can be obtained through θ_Y^* in (6), with $u_Y(\cdot)$ as in (7). Determination of θ_Y^* for the IEEE 802.11 On/Off setting specializes as follows: If the effective domain D_V of $u_V(\cdot)$ is closed from above, i.e., $\theta_V^u \triangleq \sup D_V \in D_V$ (in which case also $\theta_V^u = \sup D_Y \in D_Y$) and if, additionally, $u_V(\theta_V^u) \leq -u_C(-\theta_V^u)$ (this condition being equivalent to $f(-\theta_V^u, -u_V(\theta_V^u)) \leq 0$, because $f(\theta, \cdot)$ in (24) is a decreasing function for any θ), then $\theta_Y^* = \theta_V^u$ and the decay rate of the delay tail-probabilities $\xi_Z^* = -u_C(-\theta_V^u)$ is obtained by (24), as the unique solution in ξ of $f(-\theta_V^u, -\xi) = 0$. The case just discussed may arise when the traffic has characteristics similar to those of the example in Appendix D. In all other cases it is guaranteed that $\theta_Y^* \in D_Y^o$, so, by Item 3 of Lemma 1, θ_Y^* is a root of $u_Y(\cdot)$ in (7) (more precisely, the unique positive root, due to the monotonicity of $a_C(\cdot)$), so $u_C(-\theta_Y^*) = -u_V(\theta_Y^*)$. Therefore, in view of (24), θ_Y^* can be obtained as the unique positive solution in θ of $f(-\theta, -u_V(\theta)) = 0$. Computation of this solution will simultaneously also produce $u_V(\theta_Y^*) = -u_C(-\theta_Y^*) = \xi_Z^*$ without any additional computational cost.

Calculations for the admission control test are simpler. According to the results of Section 3, given the QoS specification ξ, one must first determine $\theta(\xi)$ in (16) and then check if inequality (17) holds. Since $u_C(-\theta(\xi)) = -\xi$, (24) suggests that $\theta(\xi)$ is the unique solution in θ of $f(-\theta, -\xi)$, thus

$$\theta(\xi) = \xi/\hat{r} - (\log \gamma_{\text{on}})^{-1}\big(-\log \gamma_{\text{off}}(\xi)\big)/\hat{r}. \tag{25}$$

This requires only a single evaluation of the function $\gamma_{\text{off}}(\cdot)$ at the argument ξ, keeping the computational complexity low. (In contrast, the computations for determining the asymptotic decay rate ξ_Z^* typically require repetitive evaluation of this function, in the course of some numerical zero finding method.) Moreover, when the payload of transmitted packets has a constant value P, (18) yields $(\log \gamma_{\text{on}})^{-1}(x) = \hat{r}x/P$, so (25) simplifies further to the closed form solution $\theta(\xi) = \xi/\hat{r} + \log \gamma_{\text{off}}(\xi)/P$. Note that, as long as the conditions[4] in the WLAN remain unchanged, a single evaluation of $\theta(\xi)$ suffices to enable an arbitrary number of admission control tests (17), each of them being invoked whenever the mobile station is about to engage a new traffic flow.

The function $u_C(\cdot)$ corresponding to the IEEE 802.11 On/Off model is such that the value of $\lim_{\theta\to-\infty} u_C(\theta)$ is always finite. Indeed, (24) suggests that

$$\log \gamma_{\text{off}}\big(-u_C(\theta)\big) = -\log \gamma_{\text{on}}\big(\hat{r}\theta - u_C(\theta)\big), \quad \forall \theta \in \mathbb{R}.$$

In order to maintain the left hand side finite, $-u_C(\theta) < \omega_{\text{off}}^*$, thus $\lim_{\theta\to-\infty} u_C(\theta) \geq -\omega_{\text{off}}^*$. Moreover, by Jensen's inequality $u_C(\theta) \geq \bar{r}_C\theta$, so when $\theta \to -\infty$ the

[4] Number of active stations in the WLAN and (if the measurement-assisted variant of the model is used), loading conditions at other stations.

argument of $\gamma_{\text{on}}(\cdot)$ in the right hand side $\hat{r}\theta - u_C(\theta) \leq (\hat{r} - \bar{r}_C)\theta \to -\infty$. Therefore, when $\theta \to -\infty$ the right hand side tends to $-\lim_{\omega\to-\infty} \log \gamma_{\text{on}}(\omega) = -\log \Pr\{T_{\text{on}} = 0\} = +\infty$, because the payload of a transmitted packet can never be empty. The left hand side must also approach infinity, thus necessarily $\lim_{\theta\to-\infty} u_C(\theta) = -\omega^*_{\text{off}}$.

As already remarked at the end of Section 3, the finiteness of this limit implies that the decay rate of the delay tail-probabilities cannot exceed ω^*_{off}. Thus, the admission control test discussed earlier must be preceded by the test $\xi < \omega^*_{\text{off}}$. If this test fails then the whole admission control test fails, otherwise the normal test is applied. The quantity ω^*_{off} is determined as explained in the beginning of this subsection. We note that the inherent reason why the decay rate ξ cannot exceed some finite bound with servers of the On/Off type is that, even when the traffic is arbitrarily low and packets arbitrarily small, the incoming packets may find the queue empty but they still have to wait until the server's residual Off period is finished before they can be processed.

5. Validation of the IEEE 802.11 model for delay-related QoS

We now validate the IEEE 802.11 Eff. Capacity model by comparing analytical results with simulation. In alignment with the paper's focus, we concentrate on delay-related QoS; for the effectiveness of the model in connection with loss-related performance see [12]. The simulation results were obtained with the help of the ns-2 simulator [30], using system parameter values corresponding to IEEE 802.11g, operating in Direct-Sequence Spread Spectrum (DSSS) Orthogonal Frequency-Division Multiplexing (OFDM) mode with

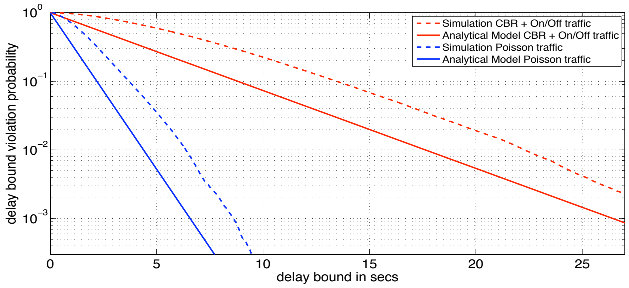

Figure 1: Delay tail-probabilities at a mobile station, in a WLAN with nine additional saturated stations, for two different traffic loads.

RTS/CTS handshaking enabled [31]. A constant payload size $P = 8184$ bits was used in all cases.

The first set of results, depicted in Fig. 1, assesses the potential of the IEEE 802.11 Eff. Capacity function to track closely the tail of the delay probabilities. A WLAN with 10 mobile stations was studied; 9 of these stations were subjected to a very high traffic load, so they operated under saturation conditions, while the 10th station was loaded with traffic of a known profile and the delay, from a packet's entrance to the station's queue until the completion of its transmission, was measured. These measurements were used for constructing the empirical complementary probability distribution function of the delay, which is plotted in the figure in semilog scale, using dashed lines.

Two simulations were run, each using a different traffic profile for the observed station: The first scenario (blue lines in Fig. 1) employed Poisson traffic with packets of constant size (equal to P), while in the second case (red lines) the traffic consisted of a superposition of constant bit rate traffic at 335.4 kbps and a bursty On/Off traffic source with exponentially distributed

On and Off periods, of mean durations equal to 0.4 sec and 0.8 sec, respectively, and a peak rate equal to 1006.2 kbps. The overall mean traffic load was the same in both scenarios, equal to 670.8 kbps.

The simulation results are accompanied by model-derived straight-line curves (solid lines in Fig. 1), at slopes equal to the corresponding asymptotic decay rates of the delay tail-probabilities, as predicted by the Eff. Bandwidth/ Capacity theory of Section 3. The relevant calculations were performed by the methodology described in Subsection 4.2, using the Eff. Bandwidth function corresponding to the traffic profile pertaining to each scenario. It can be seen that the delay tail-probabilities derived from the simulation decay exponentially, at a rate that agrees with the analytical result.

We now turn to the potential of the IEEE 802.11 function with respect to admission control decisions. A WLAN containing 10 stations is again considered, but this time all stations feature the same Poisson traffic, with packets of constant size P and 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 On/Off type, with exponentially distributed On and Off periods of mean durations equal to 0.4 sec and 0.8 sec, 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 dictates that the delay should not exceed 1 sec with probability higher than 10^{-2}. This QoS specification corresponds to a target decay rate $\xi = -\log 10^{-2}/(1\,\text{sec}) = 2\log 10\,\text{sec}^{-1}$. This, in turn, corresponds through (25), the specialized form of (16) for the IEEE 802.11 setting, to a decay rate for the queue content tail-probabilities equal to $\theta(\xi)$, which is then used in the admission control test (17). When testing for admission of the k^{th} On/Off flow, the Eff. Bandwidth function at the left hand side of (17) is set to $a_V(\cdot) = a_{\text{Poisson}}(\cdot) + ka_{\text{onoff}}(\cdot)$. It is noted that the IEEE 802.11 Eff. Capacity function used here is *not* the same as the one in the previous set of results, despite the fact that the WLAN contains 10 stations in both cases. The reason is that the other competing stations are not saturated in the present context. Thus, 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. 2, which plots probabilities of exceeding delay thresholds when the station is loaded with 4 (blue lines) and 5 (red lines) On/Off flows in addition to the Poisson background load. Results from both simulation (dashed lines) and analysis (solid lines) are included, as with Fig. 1. It may be seen that with 4 flows the probability of the delay exceeding 1 sec is below the target value 10^{-2} and that the introduction of the 5^{th} flow raises the value of this probability above the threshold, violating the QoS. Of course this was to be expected for the analytically derived straight-line curves, the result being nothing more than a manifestation of Corollary 2, as applied to the delay process $Z(t)$. However, Fig. 2 further illustrates that the exact delay probabilities (as determined by simulation) also follow the

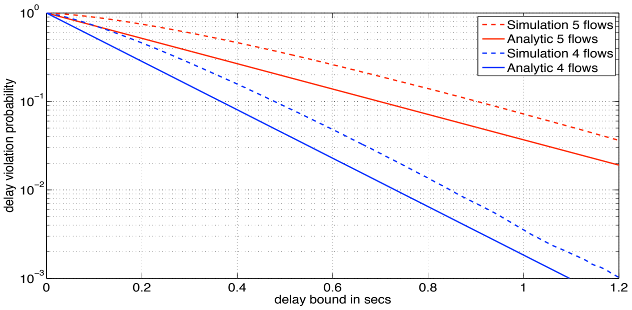

Figure 2: Delay tail-probabilities when the mobile station admits four and five On/Off flows, on top of Poisson background traffic.

predictions of the admission control test, closely enough.

6. Conclusions

The paper provided the, up to now missing, formal justification for the use of the Effective Bandwidth/Capacity theory in delay-related performance contexts. By representing the delay experienced by traffic entering a FCFS queue as the supremum of a stochastic process and by suitably extending and applying prior results, it was established rigorously that the theory is capable of providing an asymptotically tight approximation to delay tail-probabilities. In particular, the paper formalized the, previously heuristic, association of the asymptotic exponential decay rate of the queue content probabilities with its counterpart for the delay probabilities, through the server's Eff. Capacity function. The asymptotic approximation to the delay tail-probabilities was complemented by associated admission control schemes that are useful for providing delay-related QoS guarantees. The results apply to queueing systems operating in either discrete- or continuous-time and featuring arbitrary time-varying traffic and service processes, provided these processes are independent and possess well-defined Eff. Bandwidth and Eff. Capacity functions, respectively.

The general results were applied to the important setting of IEEE 802.11 WLANs, by modeling each IEEE 802.11 station as an On/Off server and then using the Effective Capacity function corresponding to this model. Computational and algorithmic details relating to the application of the general theory with the particular Eff. Capacity function of this On/Off model were also discussed. Comparison of the analytical results with simulation validated the effectiveness of the On/Off IEEE 802.11 model in providing delay-based QoS guarantees.

Appendix A. Proof of Lemma 1

Item 1 is an immediate consequence of the definition in (6). For Item 2, note that for any $0 < \theta < \theta_Y^*$ one may find a θ', such that $0 < \theta < \theta' < \theta_Y^*$, and $u_Y(\theta') \leq 0$. Thus, there exists $0 < h < 1$ such that $\theta = (1-h)0 + h\theta'$. Since $u_Y(0) = 0$, the convexity of $u_Y(\cdot)$ implies that $u_Y(\theta) \leq (1-h)u_Y(0) + hu_Y(\theta') = hu_Y(\theta') \leq 0$. To prove Item 3, note that $u_Y(\cdot)$ is convex in D_Y, hence continuous in D_Y^o. Since $\theta_Y^* \in D_Y^o$, Item 1 suggests that $u_Y(\theta_Y^*) = \lim_{\theta \downarrow \theta_Y^*} u_Y(\theta) \geq 0$, while, by Item 2, $u_Y(\theta_Y^*) = \lim_{\theta \uparrow \theta_Y^*} u_Y(\theta) \leq 0$; hence $u_Y(\theta_Y^*) = 0$. Furthermore, since $u_Y(\cdot)$ is convex and differentiable in D_Y^o, it holds

$$\frac{u_Y(\theta_2) - u_Y(\theta_1)}{\theta_2 - \theta_1} \leq u_Y'(\theta_2), \quad \forall\, \theta_2 > \theta_1. \tag{A.1}$$

By applying this result with $\theta_2 = \theta$ and $\theta_1 = \theta_Y^*$, one has

$$u_Y'(\theta) \geq \big(u_Y(\theta) - u_Y(\theta_Y^*)\big)/(\theta - \theta_Y^*) = u_Y(\theta)/(\theta - \theta_Y^*) > 0,$$

where the last inequality is due to Item 1. Now assume that there exists $\theta_o > 0$ such that $u_Y(\theta_o) < 0$. By virtue of (6), $\theta_Y^* \geq \theta_o > 0$. Also, since $u_Y(0) = u_Y(\theta_Y^*) = 0$ and $u_Y(\theta_o) < 0$, the convexity of $u_Y(\cdot)$ implies that $u_Y(\theta) < 0$ for all $\theta \in (0, \theta_o)$ and all $\theta \in (\theta_o, \theta_Y^*)$, thus there is no positive root of $u_Y(\cdot)$ in $(0, \theta_Y^*)$. By Item 1 there is no root greater than θ_Y^*, so θ_Y^* is the unique positive root of $u_Y(\cdot)$. By applying (A.1) with $\theta_1 = \theta_o$ and $\theta_2 = \theta_Y^*$, one obtains $u_Y'(\theta_Y^*) \geq -u_Y(\theta_o)/(\theta_Y^* - \theta_o) > 0$. Lastly, in connection with Item 4 assume that $\bar{r}_Y < 0$. Since $a_Y(0) = \bar{r}_Y$ and $a_Y(\cdot)$ is continuous, there exists $\theta_o > 0$ such that $a_Y(\theta_o) < 0$, so $u_Y(\theta_o) = \theta_o a_Y(\theta_o) < 0$.

Appendix B. Proof of Theorem 1

For Item 1, if $\theta_Y^* = \infty$ there is nothing to prove. Thus, assume $\theta_Y^* < \infty$. Following the reasoning of Theorem 2.1 in [15] (specialized to the linear case of interest here), for each $a > 0$,

$$\liminf_{b\to\infty} \frac{1}{b} \log \Pr\{Q > b\} \geq \liminf_{t\to\infty} \frac{1}{at} \log \Pr\{Y(t) > at\} = a^{-1} \liminf_{t\to\infty} t^{-1} \log \Pr\{Y(t)/t > a\}.$$

Let $u_Y^*(x) \triangleq \sup_{\theta\in\mathbb{R}}\{\theta x - u_Y(\theta)\}$ be the Fenchel-Legendre transform of $u_Y(\cdot)$. If $u_Y(\cdot)$ had been assumed lower semicontinuous, the 'usual' lower bound of the Gärtner-Ellis Theorem would apply and the right-hand side of the previous inequality would be bounded below by $-a^{-1} \inf_{x>a} u_Y^*(x)$, leading to the outcome of Theorem 2.1 in [15] for the special linear case.

Without the semicontinuity assumption, the weakened form of the Gärtner-Ellis lower bound applies (see, e.g., Item b of Theorem 2.3.6 in [27]), so one obtains

$$\liminf_{b\to\infty} \frac{1}{b} \log \Pr\{Q > b\} \geq -\frac{1}{a} \inf_{\substack{x>a \\ x\in\mathcal{F}}} u_Y^*(x), \quad \forall a > 0, \tag{B.1}$$

where $\mathcal{F}$ is the set of exposed points of $u_Y^*(\cdot)$ whose exposing hyperplane belongs to D_Y^o. For a complete description of an exposed point and its associated exposing hyperplane see Definition 2.3.3 in [27]. For the purposes of this proof it is sufficient to employ Item b of Lemma 2.3.9 in [27], according to which, for any $\theta \in D_Y^o$, the value $x_\theta = u_Y'(\theta)$ is an exposed point of $u_Y^*(\cdot)$ with exposing hyperplane θ and, furthermore, $u_Y^*(x_\theta) = \theta x_\theta - u_Y(\theta)$.

Now consider firstly the case $\theta_Y^* \in D_Y^o$. By Item 3 of Lemma 1, for any $\theta \in D_Y^o$ with $\theta > \theta_Y^*$, $x_\theta = u_Y'(\theta) > 0$ is a positive exposed point. Thus, for any $0 < a < x_\theta$,

$$\inf_{\substack{x>a \\ x\in\mathcal{F}}} u_Y^*(x) \leq u_Y^*(x_\theta) = \theta x_\theta - u_Y(\theta) \leq \theta x_\theta,$$

the last inequality following by Item 1 of Lemma 1, because $\theta > \theta_Y^*$. In conjunction with (B.1), one gets

$$\liminf_{b\to\infty} \frac{1}{b} \log \Pr\{Q > b\} \geq -\theta \frac{x_\theta}{a}, \quad \forall a < x_\theta,\ \theta \in D_Y^o,\ \theta > \theta_Y^*,$$

By letting $a \uparrow x_\theta$ and subsequently letting $\theta \downarrow \theta_Y^*$, the result follows.

If $\theta_Y^* \notin D_Y^o$ then one has $\theta_Y^* = \sup D_Y^o$ and, since θ_Y^* has been assumed finite, the steepness condition of Assumption 1 applies, thus

$$\lim_{\theta\uparrow\theta_Y^*} u_Y'(\theta) = +\infty. \tag{B.2}$$

Hence, for all $\theta \in D_Y^o$ suitably close to θ_Y^*, $x_\theta = u_Y'(\theta) > 0$ is a positive exposed point. Therefore,

$$-a^{-1} \inf_{\substack{x>a \\ x\in\mathcal{F}}} u_Y^*(x) \geq -a^{-1} u_Y^*(x_\theta) = -\theta \frac{x_\theta}{a} + \frac{u_Y(\theta)}{a}, \quad \forall a < x_\theta,$$

and by combining with (B.1) and letting $a \uparrow x_\theta$,

$$\liminf_{b\to\infty} \frac{1}{b} \log \Pr\{Q > b\} \geq -\theta + \frac{u_Y(\theta)}{x_\theta} = -\theta + \frac{u_Y(\theta)}{u_Y'(\theta)} \tag{B.3}$$

for all $\theta \in D_Y^o$ suitably close to θ_Y^*. By (B.2) $u_Y(\theta)$ is increasing in a neighborhood of θ_Y^*, thus $\lim_{\theta \uparrow \theta_Y^*} u_Y(\theta)$ exists and is finite (due to Item 2 of Lemma 1). Given these properties, the result follows by letting $\theta \uparrow \theta_Y^*$ in the right-hand side of (B.3). The second claim of Item 1 is immediate upon realizing that always $\limsup_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \leq 0$.

We now turn to the proof of Item 2: We want to apply Theorem 2.2 in [15] and obtain

$$I_u \triangleq \limsup_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \leq - \inf_{x>0} \frac{u_Y^*(x)}{x}. \tag{B.4}$$

Indeed, if $\mathbb{T} = \mathbb{Z}_+^o$, or if $\mathbb{T} = \mathbb{R}_+^o$ and the process $Y(t)$ additionally satisfies Hypothesis 2.3 in [15], then Theorem 2.2 in [15] is directly seen to apply. (Note that lower semicontinuity of $u_Y(\cdot)$ is not required, because the theorem uses just the upper bound of the Gärtner-Ellis Theorem, for which Assumption 1—in fact the first two items therein—suffices.) Furthermore, we will show later that (B.4) also applies when $\mathbb{T} = \mathbb{R}_+^o$ and the alternative condition in the statement of Theorem 1 holds.

For any $0 < \theta < \theta_Y^*$ (such θ exists, because $\theta_Y^* > 0$) one has $u_Y^*(x) \triangleq \sup_{\theta' \in \mathbb{R}}\{\theta' x - u_Y(\theta')\} \geq \theta x - u_Y(\theta) \geq \theta x$, where the last inequality follows from Item 2 of Lemma 1. Therefore, $u_Y^*(x)/x \geq \theta$ for any $x > 0$, so $\inf_{x>0}\{u_Y^*(x)/x\} \geq \theta$ and (B.4) leads to $I_u \leq -\theta$. By letting $\theta \uparrow \theta_Y^*$ and combing with Item 1 of this theorem, we are led to the result.

It remains to show that (B.4) holds when $\mathbb{T} = \mathbb{R}_+^o$ and $Y(t) = V(t) - C(t)$, where the processes in the difference are independent and each of them has nonnegative increments and an asymptotic cumulant generator satisfying Assumption 1. Towards this end, for any $\kappa > 0$ define $\hat{Y}_{\kappa,n} \triangleq \sup_{n\kappa \leq t < (n+1)\kappa} Y(t)$. Then, $Q = \sup_{t\geq 0} Y(t) = \sup_{n\geq 0} \hat{Y}_{\kappa,n}$. Moreover, by the nonnegativity of increments, $\hat{Y}_{n,\kappa} \le \bar{Y}_{n,\kappa} \triangleq V\big((n+1)\kappa\big) - C(n\kappa)$, so

$$I_u = \limsup_{b\to\infty} b^{-1} \log \Pr\{Q > b\} \le \limsup_{b\to\infty} b^{-1} \log \Pr\{\sup_{n\ge 0} \bar{Y}_{n,\kappa} > b\}. \qquad \text{(B.5)}$$

The discrete-time process $\bar{Y}_{n,\kappa}$ has an asymptotic cumulant generator too. Indeed,

$$\begin{aligned} u_{\bar{Y}}(\theta) &\triangleq \lim_{n\to\infty} n^{-1} \log \mathrm{E}\left[e^{\theta \bar{Y}_{n,\kappa}}\right] = \lim_{n\to\infty} n^{-1} \log \mathrm{E}\left[e^{\theta V\left((n+1)\kappa\right) - \theta C(n\kappa)}\right] \\ &= \kappa\big(u_V(\theta) + u_C(-\theta)\big) = \kappa u_Y(\theta), \qquad \text{(B.6)} \end{aligned}$$

using the independence of the two processes and (7). It follows that $u_{\bar{Y}}(\cdot)$ fulfills the conditions of Assumption 1 (and the existence of a $\theta_o > 0$ satisfying $u_{\bar{Y}}(\theta_o) < 0$), because $u_Y(\cdot)$ does. Thus, Theorem 2.2 in [15] applies to the discrete-time process $\bar{Y}_{n,\kappa}$ and bounds the right hand side of (B.5), yielding $I_u \le -\inf_{x>0}(u^*_{\bar{Y}}(x)/x)$. Moreover, (B.6) implies that $u^*_{\bar{Y}}(x) \triangleq \sup_{\theta\in\mathbb{R}}\{\theta x - u_{\bar{Y}}(\theta)\} = \kappa u^*_Y(x/\kappa)$. Therefore, $\inf_{x>0} x^{-1} u^*_{\bar{Y}}(x) = \inf_{x>0} x^{-1} u^*_Y(x)$ and (B.4) is seen to hold.

Appendix C. The steepness of $u_Z(\cdot)$

By (13), D^o_Z has a boundary point from below only if $\theta^\ell_V > -\theta^u_C$, in which case the middle branch in (12) applies as $\xi \downarrow \xi^\ell_Z$. By employing the strictly increasing function $-u_C^{-1}(-\cdot)$ in the continuous transformation $\theta = -u_C^{-1}(-\xi)$ one obtains

$$\begin{aligned} \lim_{\xi\downarrow\xi^\ell_Z} u'_Z(\xi) &= \lim_{\xi\downarrow -u_C(-\theta^\ell_V)} u'_V\big(-u_C^{-1}(-\xi)\big)/u'_C\big(u_C^{-1}(-\xi)\big) - 1 \\ &= \lim_{\theta\downarrow\theta^\ell_V} (u'_V(\theta)/u'_C(-\theta)) - 1 = +\infty, \end{aligned}$$

by the steepness of $u_V(\cdot)$ and the fact that $u'_C(-\theta^\ell_V) > 0$ and bounded (because $0 < -\theta^\ell_V < \theta^u_C$). The reasoning for the upper boundary of D^o_Z is similar:

When $\theta_V^u < -\hat{\theta}_C^\ell$, then $\lim_{\xi \uparrow -u_C(-\theta_V^u)} u_Z'(\xi) = \lim_{\theta \uparrow \theta_V^u} (u_V'(\theta)/u_C'(-\theta)) - 1 = +\infty$, for the same reasons as before. Finally, in the complementary case $-\hat{\theta}_C^\ell \leq \theta_V^u$, one has $\lim_{\xi \uparrow -u_C(\hat{\theta}_C^\ell)} u_Z'(\xi) = \lim_{\theta \uparrow -\hat{\theta}_C^\ell} (u_V'(\theta)/u_C'(-\theta)) - 1 = +\infty$, because $u_V'(-\hat{\theta}_C^\ell) > 0$ and $+\infty = \lim_{\xi \uparrow -u_C(\hat{\theta}_C^\ell)} u_T'(\xi) = \lim_{\theta \uparrow -\hat{\theta}_C^\ell} (1/u_C'(-\theta))$, due to the steepness of $u_T(\cdot)$.

Appendix D. An example of a system featuring $\boldsymbol{\theta_Y^*} \not\in \boldsymbol{D_Y^o}$ and $\boldsymbol{u_Y(\theta_Y^*) \neq 0}$

Consider a constant server rate c and Poisson arrivals of packets, whose size features a distribution with corresponding moment generator $\gamma(\theta)$, finite in $(-\infty, \theta_o]$ and infinite otherwise. Such a distribution may result from a density of the form $f(x) = \alpha s e^{-xs}/(1 + (xs)^2)$, $x \geq 0$, where α is the normalization constant. Then $\theta_o = s$ and $\gamma(\theta_o) = \alpha\pi/2$. In this example, $u_V(\theta) = \lambda(\gamma(\theta) - 1)$ with $D_V = D_Y = (-\infty, \theta_o]$. If the Poisson rate $\lambda < c\theta_o/(\gamma(\theta_o) - 1)$, then $\theta_Y^* = \theta_o$ and $\xi_Z^* = c\theta_o$, while $u_Y(\theta_o) = \lambda(\gamma(\theta_o) - 1) - c\theta_o < 0$.

References

[1] J. F. C. Kingman, A martingale inequality in the theory of queues, Proc. Camb. Phil. Soc. 60 (1964) 359–361.

[2] S. M. Ross, Bounds on the delay distribution in GI/G/1 queues, J. Appl. Prob. 11 (1974) 417–421.

[3] Z. Liu, P. Nain, D. Towsley, On a generalization of Kingman's bounds, Math. Meth. Oper. Res. 49 (2) (1999) 325–333.

[4] F. P. 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.

[5] C. S. Chang, J. A. Thomas, Effective bandwidth in high-speed digital networks, IEEE JSAC 13 (6) (1995) 1091–1100.

[6] C. S. Chang, T. Zajic, Effective bandwiths of departure processes from queues with time varying capacities, in: Proc. IEEE INFOCOM, 1995, pp. 1001–1009.

[7] 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.

[8] X. Zhang, J. Tang, H. 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.

[9] J. Tang, X. Zhang, Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks, IEEE Trans. Wireless Commun. 7 (6) (2008) 2318–2328.

[10] S. Vassilaras, A cross-layer optimized adaptive modulation and coding scheme for transmission of streaming media over wireless links, Wireless Networks 16 (4) (2010) 903–914.

[11] 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.

[12] E. Kafetzakis, K. Kontovasilis, I. Stavrakakis, A novel effective capacity-based framework for providing statistical QoS guarantees in IEEE 802.11 WLANs, Tech. rep., NCSR "Demokritos", submitted for publication, `http://www.di.uoa.gr/~mkafetz/My_files/DEMO-effcap-wlan.pdf.` (2007).

[13] C. S. Chang, Stability, queue length, and delay of deterministic and stochastic queueing networks, IEEE Trans. Autom. Control 39 (5) (1994) 913–931.

[14] P. Glynn, W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, J. Appl. Probab. 31A (1994) 131–156.

[15] N. G. Duffield, N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Math. Proc. Cambridge Phil. Soc. 118 (1995) 363–374.

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

[17] P. Glynn, W. Whitt, Large deviations behavior of counting processes and their inverses, Queueing Syst. 17 (1994) 107–128.

[18] G. Bianchi, I. Tinnirello, Remarks on IEEE 802.11 DCF performance analysis, IEEE Commun. Lett. 9 (8) (2005) 765–767.

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

[20] P. Chatzimisios, A. Boucouvalas, V. Vitsas, Packet delay analysis of the IEEE 802.11 MAC protocol, IEE Electronics Letters 18 (39) (2003) 1358–1359.

[21] M. Carvalho, J. J. Garcia-Luna-Aceves, Delay analysis of IEEE 802.11 in single-hop networks, in: Proc. IEEE ICNP '03, Atlanta, USA, 2003, pp. 146–155.

[22] A. Zanella, F. Pellegrini, Statistical characterization of the service time in saturated IEEE 802.11 networks, IEEE Commun. Lett. 9 (3) (2005) 225–227.

[23] O. Tickoo, B. Sikdar, Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks, in: Proc. IEEE INFOCOM, 2004, pp. 1404–1413.

[24] 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.

[25] 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.

[26] H. Vu, T. Sakurai, Accurate delay distribution for IEEE 802.11 DCF, IEEE Commun. Lett. 10 (4) (2006) 317–319.

[27] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, 2nd Edition, Vol. 38 of Applications of Mathematics, Stochastic Modeling and Applied Probability, Springer, 1998.

[28] 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.

[29] G. Bianchi, Performance analysis of the IEEE 802.11 distributed coordination function, IEEE JSAC 18 (3) (2000) 535–547.

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

[31] 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 (2003).