Effective Capacity: A Wireless Link Model for Support of Quality of Service

Dapeng Wu   Rohit Negi[*]

Abstract

To facilitate the efficient support of quality of service (QoS) in next-generation wireless networks, it is essential to model a wireless channel in terms of connection-level QoS metrics such as data rate, delay and delay-violation probability. However, the existing wireless channel models, *i.e.*, physical-layer channel models, do not explicitly characterize a wireless channel in terms of these QoS metrics. In this paper, we propose and develop a link-layer channel model termed *effective capacity* (EC). In this approach, we first model a wireless link by two EC functions, namely, the probability of non-empty buffer, and the QoS exponent of a connection. Then, we propose a simple and efficient algorithm to estimate these EC functions. The physical-layer analogs of these two link-layer EC functions are the marginal distribution (*e.g.*, Rayleigh/Ricean distribution) and the Doppler spectrum, respectively. The key advantages of the EC link-layer modeling and estimation are (1) ease of translation into QoS guarantees, such as delay bounds, (2) simplicity of implementation, (3) accuracy, and hence, efficiency in admission control and resource reservation. We illustrate the advantage of our approach with a set of simulation experiments, which show that the actual QoS metric is closely approximated by the QoS metric predicted by the EC link-layer model, under a wide range of conditions.

Key Words: Wireless channel model, QoS, delay, Doppler spectrum, fading, queueing theory.

[*] This work was supported in part by the National Science Foundation under the grant ANI-0111818. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. Please direct all correspondence to Dapeng Wu, Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Email: dpwu@cs.cmu.edu. URL: http://www.cs.cmu.edu/˜dpwu. Rohit Negi is with Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Email: negi@ece.cmu.edu. URL: http://www.ece.cmu.edu/˜negi.

1 Introduction

The next-generation wireless networks such as the third generation (3G) and the fourth generation (4G) wireless systems are targeted at supporting diverse quality of service (QoS) requirements and traffic characteristics [9]. The success in the deployment of such networks will critically depend upon how efficiently the wireless networks can support traffic flows with QoS guarantees [10]. To achieve this goal, mechanisms for guaranteeing QoS (*e.g.*, admission control and resource reservation) need to be efficient and practical [6].

Efficient and practical mechanisms for QoS support require accurate and simple channel models [10]. Towards this end, it is essential to model a wireless channel in terms of QoS metrics such as data rate, delay and delay-violation probability. However, the existing channel models (*e.g.*, Rayleigh fading model with a specified Doppler spectrum) do not explicitly characterize a wireless channel in terms of these QoS metrics. To use the existing channel models for QoS support, we first need to estimate the parameters for the channel model, and then extract QoS metrics from the model. This two-step approach is obviously complex, and may lead to inaccuracies due to possible approximations in extracting QoS metrics from the models.

To address this issue, we propose and develop a link-layer channel model termed the *effective capacity* (EC) model. In this approach, we first model a wireless link by two EC functions, namely, the probability of non-empty buffer, and the QoS exponent of the connection. Then, we propose a simple and efficient algorithm to estimate these EC functions. The physical-layer analogs of these two link-layer EC functions are the marginal distribution (*e.g.*, Rayleigh/Ricean distribution) and the Doppler spectrum, respectively. The key advantages of EC link-layer modeling and estimation are (1) ease of translation into QoS guarantees, such as delay bounds, (2) simplicity of implementation, (3) accuracy, and hence, efficiency in admission control and resource reservation. Simulation results show that the actual QoS metric is closely approximated by the estimated QoS metric obtained from our channel estimation algorithm, under a wide range of conditions. This demonstrates the effectiveness of the EC link-layer model, in guaranteeing QoS.

Conventional channel models directly characterize the fluctuations in the amplitude of a radio signal. We call these models *physical-layer channel* models, to distinguish them from the *link-layer channel* model we propose. In this paper, we consider small-scale fading model [12] for the physical-layer channel. Small-scale fading models describe the characteristics of generic radio paths in a statistical fashion. Small-scale fading refers to the dramatic changes in signal amplitude and phase that can be experienced as a result of small changes (as small as a half-wavelength) in the spatial separation between a receiver and a transmitter. Small-scale fading can be slow or fast, depending on the Doppler spread. The statistical time-varying nature of the envelope of a flat-fading signal is characterized by distributions such as Rayleigh, Ricean, Nakagami, etc. [12].

Physical-layer channel models provide a quick estimate of the physical-layer performance of wireless communications systems (*e.g.*, symbol error rate vs. signal-to-noise ratio (SNR)). However, physical-layer channel models cannot be easily translated into complex link-layer QoS guarantees for a connection, such as bounds on delay. The reason is that, these complex QoS requirements

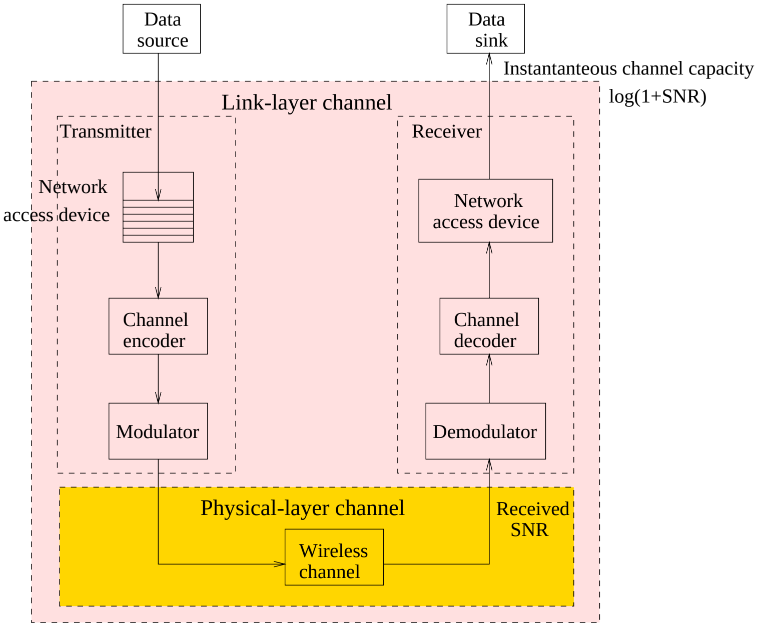


Figure 1: A packet-based wireless communication system.

need an analysis of the queueing behavior of the connection, which is hard to extract from physical-layer models. Thus it is hard to use physical-layer models in QoS support mechanisms, such as admission control and resource reservation.

Recognizing that the limitation of physical-layer channel models in QoS support, is the difficulty in analyzing queues using them, we propose moving the channel model up the protocol stack, from the physical-layer to the link-layer. We call the resulting model an *effective capacity* link model, because it captures a generalized link-level capacity notion of the fading channel. Figure 1 illustrates the difference between the conventional physical-layer and our proposed link-layer model.[1] For simplicity, we interchange "physical-layer channel" with "physical channel" and interchange "link-layer channel" with "link" in the rest of the paper.

To summarize, the effective capacity link model that we propose, aims to characterize wireless channels in terms of functions that can be easily mapped to link-level QoS metrics, such as delay-bound violation probability. Furthermore, we propose a novel channel estimation algorithm that allows practical and accurate measurements of the effective capacity model functions.

The remainder of this paper is organized as follows. In Section 2, we elaborate on the QoS guarantees that motivate us to search for a link-layer model. We describe usage parameter control (UPC) traffic characterization, and its dual, the service curve (SC) network service characterization. We show that these concepts, borrowed from networking literature, lead us to consider the effective capacity model of wireless channels. In Section 3, we formally define the effective capacity

[1] In Figure 1, we use Shannon's channel capacity to represent the instantaneous channel capacity. In practical situations, the instantaneous channel capacity is $\log(1 + SNR/\Gamma_{link})$, where Γ_{link} is determined by the modulation scheme and the channel code used.

link model, in terms of two functions, probability of non-empty buffer and QoS exponent. We then describe an estimation algorithm, which accurately estimates these functions, with very low complexity. Section 4 shows simulation results that demonstrate the advantage of using the EC link model to accurately predict QoS, under a variety of conditions. This leads to efficient admission control and resource reservation. Section 5 concludes this paper and points out future research directions. Table 1 lists the notations used in this paper.

2 Motivation for Using Link-layer Channel Models

Physical-layer channel models have been extremely successful in wireless transmitter/receiver design, since they can be used to predict physical-layer performance characteristics such as bit/frame error rates as a function of SNR. These are very useful for circuit switched applications, such as cellular telephony. However, future wireless systems will need to handle increasingly diverse multimedia traffic, which are expected to be primarily packet switched. For example, the new Wideband Code Division Multiple Access (W-CDMA) specifications make explicit provisions for 3G networks to evolve over time, from circuit switching to packet switching. The key difference between circuit switching and packet switching, from a link-layer design viewpoint, is that packet switching requires *queueing* analysis of the link. Thus, it becomes important to characterize the effect of the data traffic pattern, as well as the channel behavior, on the performance of the communication system.

QoS guarantees have been heavily researched in the *wired* networks (*e.g.*, Asynchronous Transfer Mode (ATM) and Internet Protocol (IP) networks). These guarantees rely on the queueing model shown in Figure 2. This figure shows that the source traffic and the network service are matched using a First-In-First-Out (FIFO) buffer (queue). Thus, the queue prevents loss of packets that could occur when the source rate is more than the service rate, at the expense of increasing the delay. Queueing analysis, which is needed to design appropriate admission control and resource reservation algorithms [1, 13], requires source *traffic characterization* and *service characterization*. The most widely used approach for traffic characterization, is to require that the amount of data (*i.e.*, bits as a function of time t) produced by a source conform to an upper bound, called the *traffic envelope* $\Gamma(t)$. The service characterization for guaranteed service is a guarantee of a minimum service (*i.e.*, bits communicated as a function of time) level, specified by a *service curve* $\Psi(t)$ [7]. Functions $\Gamma(t)$ and $\Psi(t)$ are specified in terms of certain traffic and service parameters respectively. Examples include the UPC parameters used in ATM [1] for traffic characterization, and the traffic specification T-SPEC and the service specification R-SPEC fields used with the resource reservation protocol (RSVP) [2, 7] in IP networks.

To elaborate on this point, a traffic envelope $\Gamma(t)$ characterizes the source behavior in the following manner: over any window of size t, the amount of actual source traffic $A(t)$ does not exceed $\Gamma(t)$ (see Figure 3). For example, the UPC parameters specify $\Gamma(t)$ by,

$$\Gamma(t) = \min\{\lambda_p^{(s)} t, \lambda_s^{(s)} t + \sigma^{(s)}\} \tag{1}$$

where $\lambda_p^{(s)}$ is the peak data rate, $\lambda_s^{(s)}$ the sustainable rate, and $\sigma^{(s)}$ the leaky-bucket size [7]. As

Table 1: Notations.

$Pr\{\cdot\}$	:	probability of the event $\{\cdot\}$.
$\Gamma(t)$	:	a traffic envelope.
$\Psi(t)$	:	a network service curve.
$A(t)$	:	the amount of source data over the time interval $[0,\, t)$.
$S(t)$	:	the actual service of a channel in bits, over the time interval $[0,\, t)$.
$r(t)$	:	the instantaneous capacity of a channel at time t.
$\tilde{S}(t)$	:	the service provided by a channel, *i.e.*, $S(t) = \int_0^t r(\tau)\mathrm{d}\tau$.
$\lambda_p^{(s)}$	:	the peak rate of a source.
$\lambda_s^{(s)}$	:	the sustainable rate of a source.
$\sigma^{(s)}$	:	the leaky-bucket size for the source model.
$\lambda_s^{(c)}$	:	the channel sustainable rate.
$\sigma(c)$	:	the maximum fade duration of a channel.
r	:	the service rate of a queue.
B	:	the buffer size of a queue.
$\Lambda(u)$	:	the asymptotic log-moment generating function of a stochastic process.
$\alpha(u)$	:	the effective bandwidth of a source.
$\alpha^{(c)}(u)$	:	the effective capacity of a channel.
$Q(t)$	:	the length of a queue at time t.
$D(t)$	:	the delay experienced by a packet arriving at time t.
D_{max}	:	the delay bound required by a connection.
ε	:	the target QoS violation probability for a connection.
θ	:	the QoS exponent of a connection.
γ	:	probability of the event that a queue is non-empty.
$S(f)$	:	the Doppler spectrum (power spectral density) of a channel.
f_m	:	the maximum Doppler frequency for a mobile terminal.
f_c	:	the carrier frequency.
$det(.)$	:	the determinant of a matrix.
x_n	:	the nth channel gain (normalized by the noise variance).
r_{awgn}	:	the capacity of an additive white Gaussian noise (AWGN) channel.

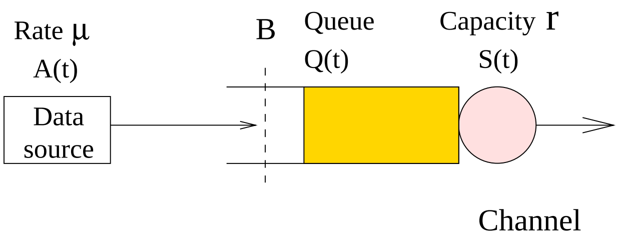


Figure 2: A queueing system model.

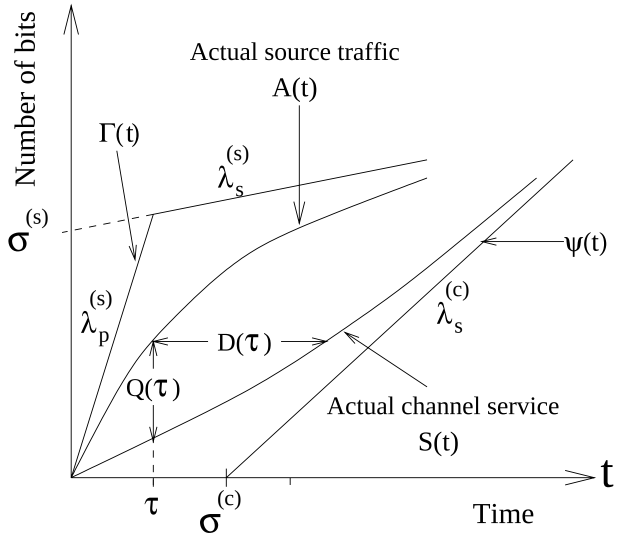


Figure 3: Traffic and service characterization.

shown in Figure 3, the curve $\Gamma(t)$ consists of two segments; the first segment has a slope equal to the peak source data rate $\lambda_p^{(s)}$, while the second segment, has a slope equal to the sustainable rate $\lambda_s^{(s)}$, with $\lambda_s^{(s)} < \lambda_p^{(s)}$. $\sigma^{(s)}$ is the Y-axis intercept of the second segment. $\Gamma(t)$ has the property that $A(t) \leq \Gamma(t)$ for any time t.

Just as $\Gamma(t)$ upper bounds the source traffic, a network service curve $\Psi(t)$ lower bounds the actual service $S(t)$ that a source will receive. $\Psi(t)$ has the property that $\Psi(t) \leq S(t)$ for any time t. Both $\Gamma(t)$ and $\Psi(t)$ are negotiated during the admission control and resource reservation phase. An example of a network service curve is the R-SPEC curve used for guaranteed service in IP networks,

$$\Psi(t) = [\lambda_s^{(c)}(t - \sigma^{(c)})]^+ \tag{2}$$

where $[x]^+ = \max\{x, 0\}$, $\lambda_s^{(c)}$ is the constant service rate and $\sigma^{(c)}$ the delay error term (due to propagation delay, link sharing and so on). This curve is illustrated in Figure 3. $\Psi(t)$ consists of two segments; the horizontal segment indicates that no packet is being serviced due to propagation delay, etc., for a time interval equal to the delay error term $\sigma^{(c)}$, while the second segment has a slope equal to the service rate $\lambda_s^{(c)}$. In the figure, we also observe that (1) the horizontal difference between $A(t)$ and $S(t)$, denoted by $D(\tau)$, is the delay experienced by a packet arriving at time τ; (2) the vertical difference between the two curves, denoted by $Q(\tau)$, is the queue length built up at time τ, due to packets that have not been served yet.

In contrast to packet-switched wireline networks, providing QoS guarantees in packet-switched wireless networks is a challenging problem. This is because wireless channels have low reliability, and time varying capacities, which may cause severe QoS violations. Unlike wireline links, which typically have a constant capacity, the capacity of a wireless channel depends upon such random factors as multipath fading, co-channel interference, and noise disturbances. Consequently, providing QoS guarantees over wireless channels requires accurate models of their *time-varying capacity*, and effective utilization of these models for QoS support.

The simplicity of the service curves discussed earlier motivates us to define the time-varying capacity of a wireless channel as in (2). Specifically, we hope to lower bound the channel service using two parameters, the channel sustainable rate $\lambda_s^{(c)}$, and the maximum fade duration $\sigma^{(c)}$.[2] However, physical-layer wireless channel models do not explicitly characterize the channel in terms of such link-layer QoS metrics as data rate, delay and delay-violation probability. For this reason, we are forced to look for alternative channel models.

A tricky issue that surfaces, is that a wireless channel has a capacity that varies *randomly* with time. Thus, an attempt to provide a strict lower bound (*i.e.*, the deterministic service curve $\Psi(t)$, used in IP networks) will most likely result in extremely conservative guarantees. For example, in a Rayleigh or Ricean fading channel, the only lower bound that can be *deterministically* guaranteed is a capacity[3] of zero! This conservative guarantee is clearly useless. Therefore, we propose to extend the concept of deterministic service curve $\Psi(t)$, to a *statistical* version, specified as the pair $\{\Psi(t), \varepsilon\}$. The statistical service curve $\{\Psi(t), \varepsilon\}$ specifies that the service provided by the channel, denoted as $\tilde{S}(t)$, will always satisfy the property that $\sup_t Pr\{\tilde{S}(t) < \Psi(t)\} \leq \varepsilon$. In other words, ε is the probability that the wireless channel will not be able to support the pledged service curve $\Psi(t)$. For most practical values of ε, a *non-zero* service curve $\Psi(t)$ can be guaranteed.

To summarize, we propose to extend the QoS mechanisms used in wired networks to wireless links, by using the traffic and service characterizations popularly used in wired networks; namely the traffic envelope $\Gamma(t)$ and the service curve $\Psi(t)$ respectively. However, recognizing that the time-varying wireless channel cannot deterministically guarantee a useful service curve, we propose to use a statistical service curve $\{\Psi(t), \varepsilon\}$.

As mentioned earlier, it is hard to extract a statistical service curve using the existing physical-layer channel models. In fact, in Section 3.4, we show how physical-layer channel models can be used to derive $\{\Psi(t), \varepsilon\}$, in an integral form. There, the reader will see that 1) it is not always possible to extract $\{\Psi(t), \varepsilon\}$ from the physical-layer model (such as, when only the Doppler spectrum, but not the higher-order statistics are known), and 2) even if it is possible, the computation involved may make the extraction extremely hard to implement. This motivates us to consider link-layer modeling, which we describe in Section 3. The philosophy here is that, we want to model the wireless channel at the layer in which we intend to use the model.

[2] $\lambda_s^{(c)}$ and $\sigma^{(c)}$ are meant to be in a statistical sense. The maximum fade duration $\sigma^{(c)}$ is a parameter that relates the delay constraint to the channel service; it determines the probability $\sup_t Pr\{S(t) < \Psi(t)\}$. We will see later that $\sigma^{(c)}$ is specified by the source with $\sigma^{(c)} = D_{max}$, where D_{max} is the delay bound required by the source.

[3] The capacity here is meant to be delay-limited capacity, which is the maximum rate achievable with a prescribed delay bound (see [8] for details).

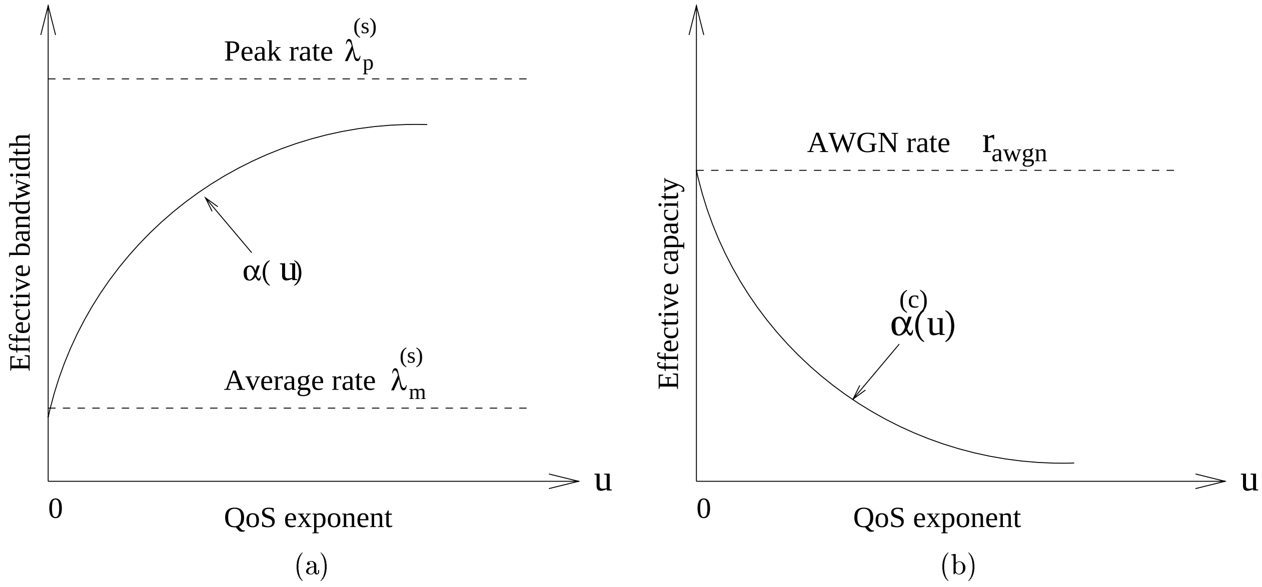


Figure 4: (a) Effective bandwidth function $\alpha(u)$ and (b) Effective capacity function $\alpha^{(c)}(u)$.

3 Effective Capacity Model of Wireless Channels

Section 2 argued that QoS guarantees can be achieved if a statistical service curve can be calculated for the given wireless link. Thus, we need to calculate a service curve $\Psi(t)$, such that for a given $\varepsilon > 0$, the following probability bound on the channel service $\tilde{S}(t)$ is satisfied,

$$\sup_t Pr\{\tilde{S}(t) < \Psi(t)\} \leq \varepsilon \tag{3}$$

Further, $\Psi(t)$ is restricted to being specified by the parameters $\{\lambda_s^{(c)}, \sigma^{(c)}\}$, as below ((2), which we reproduce, for convenience),

$$\Psi(t) = [\lambda_s^{(c)}(t - \sigma^{(c)})]^+ \tag{4}$$

Therefore, the statistical service curve specification requires that we relate its parameters $\{\lambda_s^{(c)}, \sigma^{(c)}, \varepsilon\}$ to the fading wireless channel. Note that a (non-fading) AWGN channel of capacity r_{awgn} can be specified by the triplet $\{r_{awgn}, 0, 0\}$. *i.e.*, an AWGN channel can guarantee constant data rate.

At first sight, relating $\{\lambda_s^{(c)}, \sigma^{(c)}, \varepsilon\}$ to the fading wireless channel behavior seems to be a hard problem. However, at this point, we use the idea that the service curve $\Psi(t)$ is a *dual* of the traffic envelope $\Gamma(t)$. A rich body of literature exists on the so-called *theory of effective bandwidth* [3], which models the statistical behavior of *traffic*. In particular, the theory shows that the relation

$$\sup_t Pr\{Q(t) \geq B\} \leq \varepsilon \tag{5}$$

is satisfied for large B, by choosing two parameters (which are functions of the channel rate r) that depend on the actual data traffic; namely, the probability of non-empty buffer, and the effective bandwidth of the source. *Thus, a source model defined by these two functions fully characterizes the source from a QoS viewpoint.* The duality between (3) and (5) indicates that it may be possible to adapt the theory of effective bandwidth to service curve characterization. This adaptation will point to a new channel model, which we call the *effective capacity (EC) link model.* Thus, the EC link model can be thought of as the dual of the effective bandwidth source model, which is commonly used in networking.

The rest of this section is organized as follows. In Section 3.1, we present the theory of effective bandwidth using the framework of Chang and Thomas [3]. An accurate and efficient source traffic estimation algorithm exists [11], which can be used to estimate the functions of the *effective bandwidth source model.* Therefore, we use a dual estimation algorithm to estimate the functions of the proposed *effective capacity link* model in Section 3.2. In Section 3.3, we provide physical interpretation of our link model. Section 3.4 shows that in the special case of *Rayleigh fading channel at low SNRs*, it is possible to extract the service curve from a physical-layer channel model. For Rayleigh fading channels at high SNRs, the extraction is complicated, whereas the extraction may not even be possible for other types of fading. Therefore, our link-layer EC model has substantial advantage over physical-layer models, in specifying service curves, and hence QoS.

3.1 Theory of Effective Bandwidth

The stochastic behavior of a source traffic process can be modeled asymptotically by its effective bandwidth. Consider an arrival process $\{A(t),\ t \geq 0\}$ where $A(t)$ represents the amount of source data (in bits) over the time interval $[0,\ t)$. Assume that the asymptotic log-moment generating function of $A(t)$, defined as

$$\Lambda(u) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{uA(t)}], \tag{6}$$

exists for all $u \geq 0$. Then, the *effective bandwidth function* of $A(t)$ is defined as

$$\alpha(u) = \frac{\Lambda(u)}{u} \qquad ,\ \forall\ u \geq 0. \tag{7}$$

See Ref. [3] for details.

Consider a queue of infinite buffer size served by a channel of *constant* service rate r (see Figure 2), such as an AWGN channel. Due to the possible mismatch between $A(t)$ and $S(t)$, the queue length $Q(t)$ (see Figure 3) could be non-zero. Using the theory of large deviations, it can be shown that the probability of $Q(t)$ exceeding a threshold B satisfies [3]

$$\sup_t Pr\{Q(t) \geq B\} \sim e^{-\theta_B(r)B} \qquad \text{as } B \to \infty, \tag{8}$$

where $f(x) \sim g(x)$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. However, it is found that for smaller values of B, the following approximation is more accurate [4]

$$\sup_t Pr\{Q(t) \geq B\} \approx \gamma(r) e^{-\theta_B(r) B}, \tag{9}$$

where both $\gamma(r)$ and $\theta_B(r)$ are functions of channel capacity r. According to the theory, $\gamma(r) = Pr\{Q(t) \geq 0\}$ is the *probability that the buffer is non-empty* for randomly chosen time t, while the *QoS exponent* θ_B is the solution of $\alpha(\theta_B) = r$. Thus, the pair of functions $\{\gamma(r), \theta_B(r)\}$ model the source. Note that $\theta_B(r)$ is simply the inverse function corresponding to the effective bandwidth function $\alpha(u)$.

If the quantity of interest is the delay $D(t)$ experienced by a source packet arriving at time t (see Figure 3), then the probability of $D(t)$ exceeding a delay bound D_{max} satisfies[4]

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx \gamma(r) e^{-\theta(r) D_{max}}. \tag{10}$$

Thus, the key point is that, for a source modeled by the pair $\{\gamma(r), \theta(r)\}$, which has a communication delay bound of D_{max}, and can tolerate a delay-bound violation probability of at most ε, the effective bandwidth concept shows that the constant channel capacity should be at least r, where r is the solution to $\varepsilon = \gamma(r) e^{-\theta(r) D_{max}}$. In terms of the traffic envelope $\Gamma(t)$ (Figure 3), the slope $\lambda_s^{(s)} = r$ and $\sigma^{(s)} = r D_{max}$.

Figure 4(a) shows a typical effective bandwidth function. It can be easily proved that $\alpha(0)$ is equal to the average data rate of the source, while $\alpha(\infty)$ is equal to the peak data rate. From (10), note that a source that has a more stringent QoS requirement (*i.e.*, smaller D_{max} or smaller ε), will need a larger QoS exponent θ^*.

Ref. [11] shows a simple and efficient algorithm to estimate the source model functions $\gamma(r)$ and $\theta(r)$. In the following section, we use the duality between traffic modeling ($\{\gamma(r), \theta(r)\}$), and channel modeling to propose an effective capacity link model, specified by a pair of functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$. It is clear that we intend $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ to be the channel duals of the source functions $\{\gamma(r), \theta(r)\}$. Just as the constant *channel rate* r is used in source traffic modeling, we use the constant *source traffic rate* μ in modeling the channel. Furthermore, we adapt the source estimation algorithm in [11] to estimate the link model functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$.

3.2 Effective Capacity Link Model

Let $r(t)$ be the instantaneous channel capacity at time t. Define $\tilde{S}(t) = \int_0^t r(\tau) \mathrm{d}\tau$, which is the service provided by the channel. Note that the channel service $\tilde{S}(t)$ is different from the actual service $S(t)$ received by the source; $\tilde{S}(t)$ only depends on the instantaneous channel capacity and thus is independent of the arrival $A(t)$. Paralleling the development in Section 3.1, we assume that,

$$\Lambda^{(c)}(-u) = \lim_{t \to \infty} \frac{1}{t} \log E[e^{-u\tilde{S}(t)}] \tag{11}$$

[4] $\theta(r)$ in (10) is different from $\theta_B(r)$ in (9). The relationship between them is $\theta(r) = \theta_B(r) \times r$ [15, page 57].

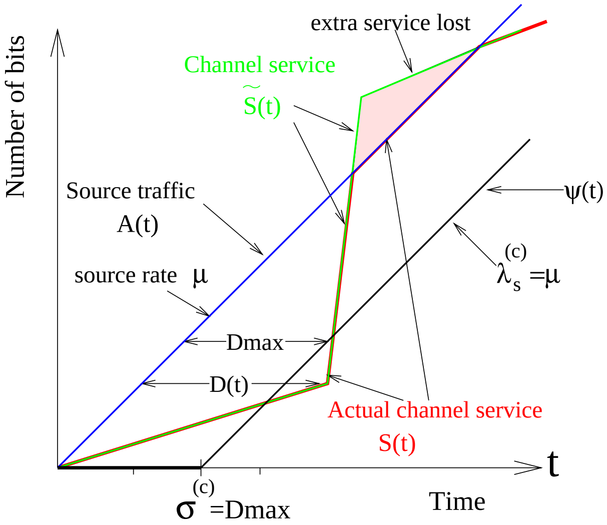


Figure 5: Relation between service curve and delay-bound violation.

exists for all $u \geq 0$. This assumption is valid, for example, for a stationary Markov fading process $r(t)$. Then, the *effective capacity function* of $r(t)$ is defined as

$$\alpha^{(c)}(u) = \frac{-\Lambda^{(c)}(-u)}{u} \qquad , \; \forall \; u \geq 0. \tag{12}$$

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate μ (see Figure 2). The theory of effective bandwidth presented in Section 3.1 can be easily adapted to this case. The difference is that whereas in Section 3.1, the source rate was variable while the channel capacity was constant, in this section, the source rate is constant while the channel capacity is variable. Similar to (10), it can be shown that the probability of $D(t)$ exceeding a delay bound D_{max} satisfies

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx \gamma^{(c)}(\mu) e^{-\theta^{(c)}(\mu) D_{max}}. \tag{13}$$

where $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ are functions of source rate μ. The approximation in (13) is accurate for large D_{max}, but we will show later, in the simulations, that this approximation is also accurate even for smaller values of D_{max}.

For a given source rate μ, $\gamma^{(c)}(\mu) = Pr\{Q(t) \geq 0\}$ is again the *probability that the buffer is non-empty* at a randomly chosen time t, while the *QoS exponent* $\theta^{(c)}(\mu)$ is defined as $\theta(\mu) = \mu\alpha^{-1}(\mu)$, where $\alpha^{-1}(\cdot)$ is the inverse function of $\alpha^{(c)}(u)$. Thus, the pair of functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ model the link.

So, when using a link that is modeled by the pair $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$, a source that requires a communication delay bound of D_{max}, and can tolerate a delay-bound violation probability of at most ε, needs to limit its data rate to a maximum of μ, where μ is the solution to $\varepsilon = \gamma^{(c)}(\mu)e^{-\theta^{(c)}(\mu)D_{max}}$. In terms of the service curve $\Psi(t)$ (Figure 3), the channel sustainable rate $\lambda_s^{(c)} = \mu$ and $\sigma^{(c)} = D_{max}$. It is clear that when the buffer is empty, the extra service $(\tilde{S}(t) - S(t))$ will be lost, *i.e.*, not used for data transmission (see Figure 5). So, we have $\tilde{S}(t) \geq S(t)$ for any time t. Furthermore, we know that the event $\{D(t) > D_{max}\}$ and the event $\{S(t) < \Psi(t)\}$ are the same. This can be illustrated by Figure 5: whenever the curve $S(t)$ is below $\Psi(t)$, the horizontal line $D(t)$ will cross the line $\Psi(t)$, *i.e.*, we have an event $\{D(t) > D_{max}\}$, since the horizontal distance between $A(t)$ and $\Psi(t)$ is D_{max}. Then, we have

$$\begin{aligned} \sup_t Pr\{D(t) > D_{max}\} &\overset{(a)}{=} \sup_t Pr\{S(t) < \Psi(t)\} \\ &\overset{(b)}{\geq} \sup_t Pr\{\tilde{S}(t) < \Psi(t)\} \end{aligned} \tag{14}$$

where (a) follows from the fact that $\{D(t) > D_{max}\}$ and $\{S(t) < \Psi(t)\}$ are the same event, and (b) from the fact that $\tilde{S}(t) \geq S(t)$ for any t. From (14), it can be seen that $\sup_t Pr\{D(t) > D_{max}\} \leq \varepsilon$ implies $\sup_t Pr\{\tilde{S}(t) < \Psi(t)\} \leq \varepsilon$.

Figure 4(b) shows that the effective capacity $\alpha^{(c)}(u)$ decreases with increasing QoS exponent u; that is, as the QoS requirement becomes more stringent, the source rate that a wireless channel can support with this QoS guarantee, decreases. The channel sustainable rate $\lambda_s^{(c)}$ is upper bounded by the AWGN capacity r_{awgn}, and lower bounded by the minimum rate 0. Figures 4(a) and 4(b) together illustrate the duality of the effective bandwidth source model and the effective capacity link model.

The function pair $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ defines our proposed effective capacity link model. The definition of these functions shows that the EC model is a link-layer model, because it directly characterizes the queueing behavior at the link-layer. From (13), it is clear that the QoS metric can be easily extracted from the EC link model.

Now that we have shown that QoS metric calculation is trivial, once the EC link model is known, we need to specify a simple (and hopefully, accurate) channel estimation algorithm. Such an algorithm should estimate the functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ from channel measurements, such as the measured SNR or channel capacity $r(t)$.

Let us take a moment to think about how we could use the existing physical-layer channel models to estimate $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$. An obvious fact that emerges is that if a channel model specifies only the *marginal* Probability Density Function (PDF) at any time t (such as Ricean PDF), along with the Doppler spectrum (such as Gans Doppler spectrum), which is *second order statistics*, then the model does not have enough information to calculate the effective capacity function (12)! Indeed, such a calculation would need higher order joint statistics, which cannot be obtained merely from the Doppler spectrum. Thus, only approximations can be made in this case. For a Rayleigh fading distribution, the joint PDF of channel gains will be complex Gaussian, and hence the Doppler spectrum is enough to calculate the effective capacity. This result is presented in

Section 3.4. However, it will be shown there, that even in the Rayleigh fading case, the calculation is complicated, and therefore, not likely to be practical.

Assume that the channel fading process $r(t)$ is stationary and ergodic. Then, a simple algorithm to estimate the functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ is the following (adapted from [5, 11]),

$$\begin{aligned} \frac{\gamma^{(c)}(\mu)}{\theta^{(c)}(\mu)} &= E[D(t)] && (15) \\ &= \tau_s(\mu) + \frac{E[Q(t)]}{\mu} \quad , \text{ and} && (16) \\ \gamma^{(c)}(\mu) &= Pr\{D(t) > 0\} && (17) \end{aligned}$$

where $\tau_s(\mu)$ is the average remaining service time of a packet being served. Note that $\tau_s(\mu)$ is zero for a fluid model (assuming infinitesimal packet size). The intuition in (15) is that, since the distribution of $D(t)$ is approximately exponential for large D (see (13)), then $E[D(t)]$ is given by (15). Now, the delay $D(t)$ is the sum of the delay incurred due to the packet already in service, and the delay in waiting for the queue $Q(t)$ to clear. This results in equation (16), using Little's theorem. Substituting $D_{max} = 0$ in (13) results in (17).

Solving (16) for $\theta^{(c)}(\mu)$, we obtain,

$$\theta^{(c)}(\mu) = \frac{\gamma^{(c)}(\mu) \times \mu}{\mu \times \tau_s(\mu) + E[Q(t)]}. \tag{18}$$

Eqs. (17) and (18) show that the functions γ and θ can be estimated by estimating $Pr\{D(t) > 0\}$, $\tau_s(\mu)$, and $E[Q(t)]$. The latter can be estimated by taking a number of samples, say N, over an interval of length T, and recording the following quantities at the nth sampling epoch: S_n the indicator of whether a packets is in service ($S_n \in \{0, 1\}$), Q_n the number of bits in the queue (excluding the packet in service), and T_n the remaining service time of the packet in service (if there is one in service). The following sample means are computed,

$$\hat{\gamma} = \frac{1}{N} \sum_{n=1}^{N} S_n, \tag{19}$$

$$\hat{q} = \frac{1}{N} \sum_{n=1}^{N} Q_n. \tag{20}$$

and

$$\hat{\tau_s} = \frac{1}{N} \sum_{n=1}^{N} T_n. \tag{21}$$

Then, from Eq. (18), we have,

$$\hat{\theta} = \frac{\hat{\gamma} \times \mu}{\mu \times \hat{\tau}_s + \hat{q}} \tag{22}$$

Eqs. (19) through (22) constitute our channel estimation algorithm, to estimate the EC link model functions $\{\gamma^{(c)}(u), \theta^{(c)}(u)\}$. They can be used to predict the QoS by approximating Eq. (13) with

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx \hat{\gamma} e^{-\hat{\theta} D_{max}}. \tag{23}$$

Furthermore, if the ultimate objective of EC link modeling is to compute an appropriate service curve $\Psi(t)$, then as mentioned earlier, given the delay bound D_{max} and the target delay-bound violation probability ε of a connection, we can find $\Psi(t) = \{\sigma^{(c)}, \lambda_s^{(c)}\}$ by, 1) setting $\sigma^{(c)} = D_{max}$, 2) solving Eq. (23) for μ and setting $\lambda_s^{(c)} = \mu$. A fast binary search procedure that estimates $\lambda_s^{(c)}$ for a given D_{max} and ε, is shown in the Appendix.

This section introduced the effective capacity link model, which is parameterized by the pair of functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$. It was shown that these functions can be easily used to derive QoS guarantees (13), such as a bound that uses $\{D_{max}, \varepsilon\}$. Furthermore, this section specified a simple and efficient algorithm ((19) through (22)) to estimate $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$, which can then be used in (13). This completes the specification of our link-layer model.

The EC link model and its application are summarized below.

EC link model:

1. $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ is the EC link model, which exists if the log-moment generating function $\Lambda^{(c)}(-u)$ in (11) exists (*e.g.*, for a stationary Markov fading process $r(t)$).

2. In addition to its stationarity, if $r(t)$ is also ergodic, then $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ can be estimated by Eqs. (19) through (22).

3. Given the EC link model, the QoS $\{\mu, D_{max}, \varepsilon\}$ can be computed by Eq. (23), where $\varepsilon = \sup_t Pr\{D(t) \geq D_{max}\}$.

4. The resulting QoS $\{\mu, D_{max}, \varepsilon\}$ corresponds directly to the service curve specification $\{\lambda_s^{(c)}, \sigma^{(c)}, \varepsilon'\}$ with $\lambda_s^{(c)} = \mu$, $\sigma^{(c)} = D_{max}$, and $\varepsilon' \leq \varepsilon$.

3.3 Physical Interpretation of Our Model $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$

We stress that the model presented in the previous section, $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$, is not just a result of mathematics (*i.e.*, large deviation theory). But rather, the model has direct physical interpretation, *i.e.*, $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ corresponds to marginal Cumulative Distribution Function (CDF) and Doppler spectrum of the underlying physical-layer channel. This correspondence can be illustrated as follows.

- The probability of non-empty buffer, $\gamma^{(c)}(\mu)$, is similar to the concept of marginal CDF (*e.g.*, Rayleigh/Ricean distribution), or equivalently, outage probability (the probability that the received SNR falls below a certain specified threshold). As shown later in Figure 9, different marginal CDF of the underlying physical-layer channel, corresponds to different $\gamma^{(c)}(\mu)$. However, the two functions, marginal CDF (*i.e.*, outage probability) and $\gamma^{(c)}(\mu)$, are not equal. The reason is that the probability of non-empty buffer takes into account the effect of packet accumulation in the buffer, while the outage probability does not (*i.e.*, an arrival packet will be immediately discarded if the SNR falls below a threshold). Therefore, the probability of non-empty buffer is larger than the outage probability, because buffering causes longer busy periods, compared with the non-buffered case.

  From Figure 9, we observe that $\gamma^{(c)}(\mu)$ and marginal CDF have similar behavior, *i.e.*, 1) both increases with the source rate μ; 2) a large outage probability at the physical layer results in a large $\gamma^{(c)}(\mu)$ at the link layer. Thus, $\gamma^{(c)}(\mu)$ does reflect the marginal CDF of the underlying wireless channel.

- $\theta^{(c)}(\mu)$, defined as the decay rate of the probability $\sup_t Pr\{D(t) \geq D_{max}\}$, corresponds to the Doppler spectrum. This can be seen from Figure 10. As shown in the figure, different Doppler rates give different $\theta^{(c)}(\mu)$. In addition, the figure shows that $\theta^{(c)}(\mu)$ increases with the Doppler rate. The reason for this is the following. As the Doppler rate increases, the time diversity of the channel also increases. This implies lower delay-violation probability $\sup_t Pr\{D(t) \geq D_{max}\}$, which leads to a larger decay rate $\theta^{(c)}(\mu)$. Therefore, $\theta^{(c)}(\mu)$ does reflect the Doppler spectrum of the underlying physical channel.

Note that a stationary Markov fading process (as is commonly assumed, for a physical wireless channel), $r(t)$, will always have a log-moment generating function $\Lambda^{(c)}(u)$. Therefore, the EC link model is applicable to such a case. On the other hand, in Section 4.2, the simulation results show that the delay-violation probability $\sup_t Pr\{D(t) > D_{max}\}$ does decrease exponentially with the delay bound D_{max} (see Figures 11 to 13). Therefore, our link model is reasonable, not only from a theoretical viewpoint (*i.e.*, Markovian property of fading channels) but also from an experimental viewpoint (*i.e.*, the actual delay-violation probability decays exponentially with the delay bound).

3.4 QoS Guarantees Using Physical-layer Channel Models

In this section, we show that in contrast to our approach, which uses a link-layer model, the existing physical-layer channel models cannot be easily used to extract QoS guarantees. In Ref. [14], the authors attempt to model the wireless physical-layer channel using a discrete state representation. For example, Ref. [14] approximates a Rayleigh flat fading channel as a multi-state Markov chain, whose states are characterized by different bit error rates. With the multi-state Markov chain model, the performance of the link layer can be analyzed, but only at expense of enormous complexity. In this section, we outline a similar method for the Rayleigh flat fading channel at low SNRs.

As mentioned earlier, $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ cannot be calculated using (12), in general, if only the marginal PDF at any time t and the Doppler spectrum are known. However, such an analytical calculation is possible for a Rayleigh flat fading channel in AWGN, albeit at very high complexity.

Suppose that the wireless channel is a Rayleigh flat fading channel in AWGN with Doppler spectrum $S(f)$. Assume that we have perfect causal knowledge of the channel gains. For example, the Doppler spectrum $S(f)$ from the Gans model [12] is the following

$$S(f) = \frac{1.5}{\pi f_m \sqrt{1 - (\frac{F}{f_m})^2}}, \tag{24}$$

where f_m is the maximum Doppler frequency; f_c is the carrier frequency; and $F = f - f_c$.

We show how to calculate the effective capacity for this channel. Denote a sequence of N measurements of the channel gain, spaced at a time-interval δ apart, by $\mathbf{x} = [x_0, x_1, \cdots, x_{N-1}]$, where $\{x_n,\ n \in [0, N-1]\}$ are the complex-valued channel gains ($|x_n|$ are therefore Rayleigh distributed). Without loss of generality, we have absorbed the constant noise variance into the definition of x_n. The measurement x_n is a realization of a random variable sequence denoted by X_n, which can be written as the vector $\mathbf{X} = [X_0, X_1, \cdots, X_{N-1}]$. The PDF of a random vector $\mathbf{X}$ for the Rayleigh fading channel is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\pi^N det(\mathbf{R})} e^{-\mathbf{x}^H \mathbf{R}^{-1} \mathbf{x}}, \tag{25}$$

where $\mathbf{R}$ is the covariance matrix of the random vector $\mathbf{X}$, $det(\mathbf{R})$ the determinant of matrix $\mathbf{R}$, and $\mathbf{x}^H$ the conjugate of $\mathbf{x}$. Now, to calculate the effective capacity, we first need to calculate,

$$\begin{aligned} E[e^{-u\tilde{S}(t)}] &= E[e^{-u\int_0^t r(\tau)d\tau}] \\ &\overset{(a)}{\approx} \int e^{-u(\sum_{n=0}^{N-1} \delta \times r(\tau_n))} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &\overset{(b)}{=} \int e^{-u(\sum_{n=0}^{N-1} \delta \log(1+|x_n|^2))} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &\overset{(c)}{=} \int e^{-u(\sum_{n=0}^{N-1} \delta \log(1+|x_n|^2))} \frac{1}{\pi^N det(\mathbf{R})} e^{-\mathbf{x}^H \mathbf{R}^{-1} \mathbf{x}} d\mathbf{x} \end{aligned} \tag{26}$$

where (a) approximates the integral by a sum, (b) from the standard result on Gaussian channel capacity (*i.e.*, $r(\tau_n) = \log(1 + |x_n|^2)$, where $|x_n|$ is the modulus of x_n), and (c) from Eq. (25). This gives the effective capacity (12) as,

$$\alpha^{(c)}(u) = \frac{-1}{u} \lim_{t \to \infty} \frac{1}{t} \log \int e^{-u(\sum_{n=0}^{N-1} \delta \log(1+|x_n|^2))} \frac{1}{\pi^N det(\mathbf{R})} e^{-\mathbf{x}^H \mathbf{R}^{-1} \mathbf{x}} d\mathbf{x} \tag{27}$$

In general, the integral in (27) is of high dimension (*i.e.*, $2N$ dimensions) and it does not reduce to a simple form, except for the case of low SNR, where approximation can be made. Next, we show a simple form of (27), for the case of low SNR. We first simplify (26) as follows.

$$
\begin{aligned}
E[e^{-u\tilde{S}(t)}] &\overset{(a)}{\approx} \int e^{-u\delta(\sum_{n=0}^{N-1}|x_n|^2)} \frac{1}{\pi^N det(\mathbf{R})} e^{-\mathbf{x}^H \mathbf{R}^{-1}\mathbf{x}} d\mathbf{x} \\
&\overset{(b)}{=} \int e^{-u\delta||\mathbf{x}||^2} \frac{1}{\pi^N det(\mathbf{R})} e^{-\mathbf{x}^H \mathbf{R}^{-1}\mathbf{x}} d\mathbf{x} \\
&\overset{(c)}{=} \frac{1}{\pi^N det(\mathbf{R})} \int e^{-\mathbf{x}^*(\mathbf{R}^{-1}+u\delta\mathbf{I})\mathbf{x}} d\mathbf{x} \\
&= \frac{1}{\pi^N det(\mathbf{R})} \times \pi^N det((\mathbf{R}^{-1}+u\delta\mathbf{I})^{-1}) \\
&= \frac{1}{det(u\delta\mathbf{R}+\mathbf{I})} \qquad (28)
\end{aligned}
$$

where (a) using the approximation $\log(1+|x_n|^2) \approx |x_n|^2$ for Eq. (26) (if $|x_n|$ is small, that is, low SNR), (b) by the definition of the norm of the vector $\mathbf{x}$, and (c) by the relation $||\mathbf{x}||^2 = \mathbf{x}^*\mathbf{x}$ (where $\mathbf{I}$ is identity matrix). We consider three cases of interest for Eq. (28):

- *Case 1 (special case):* Suppose $\mathbf{R} = r\mathbf{I}$, where $r = E|x_n|^2$ is the average channel capacity. This case happens when a mobile moves very fast with respect to the sample period. From Eq. (28), we have

  $$E[e^{-u\tilde{S}(t)}] \approx \frac{1}{det(u\delta\mathbf{R}+\mathbf{I})} = \frac{1}{(ur\delta+1)^N} \overset{(a)}{=} \frac{1}{(ur \times \frac{t}{N}+1)^N} \qquad (29)$$

  where (a) follows from the fact that the sample period δ is $\frac{t}{N}$.

  As the number of samples $N \to \infty$, we have,

  $$\lim_{N\to\infty} E[e^{-u\tilde{S}(t)}] \approx \lim_{N\to\infty} \frac{1}{(ur \times \frac{t}{N}+1)^N} = e^{-urt} \qquad (30)$$

  Thus, in the limiting case, the Rayleigh fading channel reduces to an AWGN channel. Note that this result would not apply at high SNRs, because of the concavity of the $\log(\cdot)$ function. Since Case 1 has the highest degree of diversity, it is the best case for guaranteeing QoS, *i.e.*, it provides the largest effective capacity among all the Rayleigh fading processes with the same marginal PDF. It is also the best case for high SNR.

- *Case 2 (special case):* Suppose $\mathbf{R} = r\mathbf{1}\cdot\mathbf{1}^T$, where $(.)^T$ denotes matrix transpose, and $\mathbf{1} = [1, 1, \cdots, 1]^T$. Thus, all the samples are fully correlated, which could occur if the wireless terminal is immobile. From Eq. (28), we have,

  $$E[e^{-u\tilde{S}(t)}] \approx \frac{1}{det(u\delta\mathbf{R}+\mathbf{I})} = \frac{1}{ur\delta N+1} = \frac{1}{ur \times \frac{t}{N} \times N+1} = \frac{1}{1+urt} \qquad (31)$$

  Since Case 2 has the lowest degree of diversity, it is the worst case. Specifically, Case 2 provides zero effective capacity because a wireless terminal could be in a deep fade forever, making it impossible to guarantee any non-zero capacity. It is also the worst case for high SNR.

- *Case 3 (general case):* Denote the eigenvalues of matrix $\mathbf{R}$ by $\{\lambda_n,\ n \in [0, N-1]\}$. Since $\mathbf{R}$ is symmetric, we have $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, where $\mathbf{U}$ is a unitary matrix; $\mathbf{U}^H$ is its Hermitian; and the diagonal matrix $\mathbf{\Sigma} = diag(\lambda_0, \lambda_1, \cdots, \lambda_{N-1})$. From Eq. (28), we have,

$$\begin{aligned} E[e^{-u\tilde{S}(t)}] &\approx \frac{1}{det(u\delta\mathbf{R} + \mathbf{I})} \\ &= \frac{1}{det(u\delta\mathbf{U}\mathbf{\Sigma}\mathbf{U}^H + \mathbf{U}\mathbf{U}^H)} \\ &= \frac{1}{det(\ \mathbf{U}\ diag(u\delta\lambda_0 + 1, u\delta\lambda_1 + 1, \cdots, u\delta\lambda_{N-1} + 1)\ \mathbf{U}^H\ )} \\ &= \frac{1}{\Pi_n(u\delta\lambda_n + 1)} \\ &= e^{-\sum_n \log(u\delta\lambda_n + 1)} \end{aligned} \tag{32}$$

Case 3 is the general case for a Rayleigh flat fading channel at low SNRs.

We now use the calculated $E[e^{-u\tilde{S}(t)}]$ to derive the log-moment generating function as,

$$\begin{aligned} \Lambda^{(c)}(-u) &= \lim_{t\to\infty} \frac{1}{t} \log E[e^{-u\tilde{S}(t)}] \\ &\overset{(a)}{\approx} \lim_{t\to\infty} \frac{1}{t} \log e^{-\sum_n \log(u\delta\lambda_n + 1)} \\ &\overset{(b)}{=} \lim_{\Delta f\to 0} -\Delta f \sum_n \log(u\frac{\lambda_n}{B_w} + 1) \\ &\overset{(c)}{=} -\int \log(uS(f) + 1)df \end{aligned} \tag{33}$$

where (a) follows from Eq. (32), (b) follows from the fact that the frequency interval $\Delta f = 1/t$ and the bandwidth $B_w = 1/\delta$, and (c) from the fact that the power spectral density $S(f) = \lambda_n/B_w$ and that the limit of a sum becomes an integral. This gives the effective capacity (12) as,

$$\alpha^{(c)}(u) = \frac{\int \log(uS(f) + 1)df}{u} \tag{34}$$

Thus, the Doppler spectrum allows us to calculate $\alpha^{(c)}(u)$. The effective capacity function (34) can be used to guarantee QoS using Eq. (13).

Remark 1 We argue that even if we have *perfect* knowledge about the channel gains, it is hard to extract QoS metrics from the physical-layer channel model, in the general case. The effective capacity function (34) is valid only for a Rayleigh flat fading channel, *at low SNR*. At high SNR, the

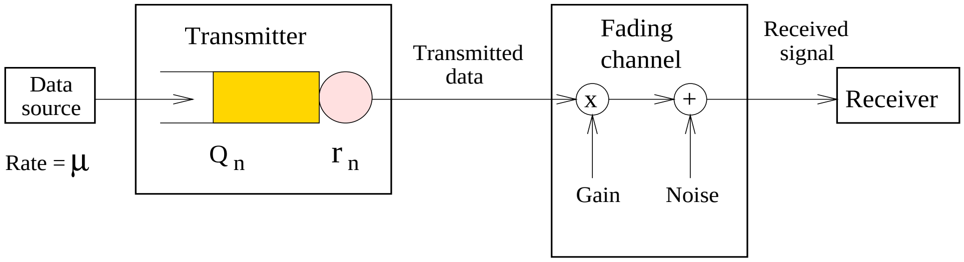


Figure 6: The queueing model used for simulations.

effective capacity for a Rayleigh fading channel is specified by the complicated integral in (27). To the best of our knowledge, a closed-form solution to (27) does not exist. It is clear that a numerical calculation of effective capacity is also very difficult, because the integral has a high dimension. Thus, it is difficult to extract QoS metrics from a physical-layer channel model, even for a Rayleigh flat fading channel. The extraction may not even be possible for more general fading channels. In contrast, the EC link model that we have proposed can be easily translated into QoS metrics for a connection, and we have shown a simple estimation algorithm to estimate the EC model functions.

4 Simulation Results

In this section, we simulate a queueing system and demonstrate the performance of our algorithm for estimating the functions of the effective capacity link model. Section 4.1 describes the simulation setting, while Section 4.2 illustrates the performance of our estimation algorithm.

4.1 Simulation Setting

We simulate the discrete-time system depicted in Figure 6. In this system, the data source generates packets at a *constant* rate μ. Generated packets are first sent to the (infinite) buffer at the transmitter, whose queue length is Q_n, where n refers to the n^{th} sample-interval. The head-of-line packet in the queue is transmitted over the fading channel at data rate r_n. The fading channel has a random channel gain x_n (the noise variance is absorbed into x_n). We use a fluid model, that is, the size of a packet is infinitesimal.

We assume that the transmitter has perfect knowledge of the channel gain x_n (the SNR, really) at each sample-interval. Therefore, it can use rate-adaptive transmissions, and strong channel coding, to transmit packets without errors. Thus, the transmission rate r_n is equal to the instantaneous (time-varying) capacity of the fading channel, as below,

$$r_n = B_c \log_2(1 + |x_n|^2) \tag{35}$$

where B_c is the channel bandwidth.

Table 2: Simulation parameters.

Channel	Maximum Doppler rate f_m	5 to 30 Hz
	AWGN channel capacity r_{awgn}	100 kb/s
	Average SNR	0/15 dB
	Sampling-interval T_s	1 ms
Source	Constant bit rate μ	30 to 85 kb/s

The average SNR is fixed in each simulation run. We define r_{awgn} as the capacity of an equivalent AWGN channel, which has the same average SNR. *i.e.*,

$$r_{awgn} = B_c \log_2(1 + SNR_{avg}) \tag{36}$$

where SNR_{avg} is the average SNR, *i.e.*, $E|x_n|^2$. Then, we can eliminate B_c using Eqs. (35) and (36) as,

$$r_n = \frac{r_{awgn} \log_2(1 + |x_n|^2)}{\log_2(1 + SNR_{avg})} \tag{37}$$

In our simulations, the sample interval is set to 1 milli-second. This is not too far from reality, since 3G WCDMA systems already incorporate rate adaptation on the order of 10 milli-second [9].

Each simulation run is 1000-second long in all scenarios. Since the channel sample rate is 1000 samples/sec, 1,000,000 samples of Rayleigh/Ricean flat fading x_n were generated for each 1000-second run, using a first-order auto-regressive (AR) model. Specifically, x_n is generated by the following AR(1) model

$$x_n = \kappa x_{n-1} + v_n, \tag{38}$$

where the noise v_n is zero-mean complex Gaussian with unit variance per dimension and is statistically independent of x_{n-1}. The coefficient κ can be determined by the following procedure: 1) compute the coherence time T_c by [12, page 165]

$$T_c \approx \frac{9}{16\pi f_m}, \tag{39}$$

where the coherence time is defined as the time, over which the time auto-correlation function of the fading process is above 0.5; 2) compute the coefficient κ by[5]

$$\kappa = 0.5^{T_s/T_c}, \tag{40}$$

where T_s is the sampling interval.

Table 2 lists the parameters used in our simulations.

[5] The auto-correlation function of the AR(1) process is κ^n, where n is the number of sample intervals. Solving $\kappa^{T_c/T_s} = 0.5$ for κ, we obtain (40).

4.2 Performance of the Estimation Algorithm

We organize this section as follows. In Section 4.2.1, we estimate the functions $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ of the effective capacity link model, from the measured x_n. Section 4.2.2 provides simulation results that demonstrate the relation between the physical channel and our link model. In Section 4.2.3, we show that the estimated EC functions accurately predict the QoS metric, under a variety of conditions.

4.2.1 Effective Capacity Model $\{\hat{\gamma}, \hat{\theta}\}$ Estimation

In the simulations, a Rayleigh flat fading channel is assumed. We simulate four cases: 1) $SNR_{avg} =$ 15 dB and the maximum Doppler rate $f_m = 5$ Hz, 2) $SNR_{avg} = 15$ dB and $f_m = 30$ Hz, 3) $SNR_{avg} = 0$ dB and $f_m = 5$ Hz, and 4) $SNR_{avg} = 0$ dB and $f_m = 30$ Hz. Figures 7 and 8 show the estimated EC functions $\hat{\gamma}(\mu)$ and $\hat{\theta}(\mu)$. As the source rate μ increases from 30 kb/s to 85 kb/s, $\hat{\gamma}(\mu)$ increases, indicating a higher buffer occupancy, while $\hat{\theta}(\mu)$ decreases, indicating a slower decay of the delay-violation probability. Thus, the delay-violation probability is expected to increase, with increasing source rate μ. From Figure 7, we also observe that SNR has a substantial impact on $\hat{\gamma}(\mu)$. This is because higher SNR results in larger channel capacity, which leads to smaller probability that a packet will be buffered, *i.e.*, smaller $\hat{\gamma}(\mu)$. In contrast, Figure 7 shows that f_m has little effect on $\hat{\gamma}(\mu)$. The reason is that $\hat{\gamma}(\mu)$ reflects the marginal CDF of the underlying fading process, rather than the Doppler spectrum.

4.2.2 Physical Interpretation of Link Model $\{\hat{\gamma}, \hat{\theta}\}$

To illustrate that different physical channel induces different parameters $\{\hat{\gamma}, \hat{\theta}\}$, we simulate two kinds of channels, *i.e.*, a Rayleigh flat fading channel and a Ricean flat fading channel. For the Rayleigh channel, we set the average SNR to 15 dB. For the Ricean channel, we set the K factor[6] to 3 dB. We simulate two scenarios: A) changing the source rate while fixing the Doppler rate at 30 Hz, and B) changing the Doppler rate while fixing the source rate, *i.e.*, $\mu = 85$ kb/s.

The result for scenario A is shown in Figure 9. For comparison, we also plot the marginal CDF (*i.e.*, Rayleigh/Ricean CDF) of the physical channel in the same figure. The marginal CDF for Rayleigh channel, *i.e.*, the probability that the SNR falls below a threshold SNR_{th}, is

$$Pr\{SNR \leq SNR_{th}\} = 1 - e^{-SNR_{th}/SNR_{avg}} \tag{41}$$

Similar to (37), we have the source rate

$$\mu = \frac{r_{awgn} \log_2(1 + SNR_{th})}{\log_2(1 + SNR_{avg})} \tag{42}$$

[6] The K factor is defined as the ratio between the deterministic signal power A^2 and the variance of the multipath $2\sigma_m^2$, *i.e.*, $K = A^2/(2\sigma_m^2)$.

Solving (42) for SNR_{th}, we obtain

$$SNR_{th} = (1 + SNR_{avg})^{\frac{\mu}{r_{awgn}}} - 1 \qquad (43)$$

Using (41) and (43), we plot the marginal CDF of the Rayleigh channel, as a function of source rate μ. Similarly, we plot the marginal CDF of the Ricean channel, as a function of source rate μ.

As shown in Figure 9, different marginal CDF at the physical layer yields different $\hat{\gamma}(\mu)$ at the link layer. We observe that $\hat{\gamma}(\mu)$ and marginal CDF have similar behavior, *i.e.*, 1) both increases with the source rate μ; 2) if one channel has a larger outage probability than another channel, it also has a larger $\hat{\gamma}(\mu)$ than the other channel. For example, in Figure 9, the Rayleigh channel has a larger outage probability and a larger $\hat{\gamma}(\mu)$ than the Ricean channel. Thus, the probability of non-empty buffer, $\hat{\gamma}(\mu)$, is similar to marginal CDF, *i.e.*, outage probability.

Figures 10 and 11 show the result for scenario B. From Figure 10, it can be seen that different Doppler rate at the physical layer leads to different $\hat{\theta}(\mu)$ at the link layer. In addition, the figure shows that $\hat{\theta}(\mu)$ increases with the Doppler rate. This is reasonable since the increase of the Doppler rate leads to the increase of time diversity, resulting in a larger decay rate $\hat{\theta}(\mu)$ of the delay-violation probability. Therefore, $\hat{\theta}(\mu)$ corresponds to the Doppler spectrum of the physical channel.

Figure 11 shows the actual delay-violation probability $\sup_t Pr\{D(t) > D_{max}\}$ vs. the delay bound D_{max}, for various Doppler rates. It can be seen that the actual delay-violation probability decreases exponentially with the delay bound D_{max}, for all the cases. This justifies the use of an exponential bound (23) in predicting QoS, thereby justifying our link model $\{\hat{\gamma}, \hat{\theta}\}$.

4.2.3 Accuracy of the QoS Metric Predicted by $\hat{\gamma}$ and $\hat{\theta}$

In the previous section, the simulation results have justified the use of $\{\hat{\gamma}, \hat{\theta}\}$ in predicting QoS. In this section, we evaluate the accuracy of such a prediction. To test the accuracy, we use $\hat{\gamma}$ and $\hat{\theta}$ to calculate the delay-bound violation probability $\sup_t Pr\{D(t) > D_{max}\}$ (using (23)), and then compare the estimated probability with the actual (*i.e., measured*) $\sup_t Pr\{D(t) > D_{max}\}$.

To show the accuracy, we simulate three scenarios. In the first scenario, the source rates μ are 75/80/85 kb/s, which loads the system as light/moderate/heavy, respectively. For all three cases, we simulate a Rayleigh flat fading channel with $SNR_{avg} = 15$ dB, $r_{awgn} = 100$ kb/s and $f_m = 30$ Hz. Figure 12(a) plots the actual and the estimated delay-bound violation probability $\sup_t Pr\{D(t) > D_{max}\}$ as a function of D_{max}. As predicted by (23), the delay-violation probability follows an exponential decrease with D_{max}. Furthermore, the estimated $\sup_t Pr\{D(t) > D_{max}\}$ is close to the actual $\sup_t Pr\{D(t) > D_{max}\}$.

In the second scenario, we also set $r_{awgn} = 100$ kb/s and $f_m = 30$ Hz, but change the average SNR to 0 dB. Figure 12(b) shows that the conclusions drawn from the first scenario still hold. Thus, our estimation algorithm gives consistent performance over different SNRs also.

In the third scenario, we set $SNR_{avg} = 15$ dB and $r_{awgn} = 100$ kb/s, but we change the Doppler rate f_m to 5 Hz. Figure 13 shows that the conclusions drawn from the first scenario still hold. Thus, our estimation algorithm consistently predicts the QoS metric under different Doppler rate f_m.

In summary, the simulations illustrate that our EC link model, together with the estimation algorithm, predict the actual QoS accurately.

5 Concluding Remarks

Efficient bandwidth allocation and QoS provisioning over wireless links, demand a simple and effective wireless channel model. In this paper, we modeled a wireless channel from the perspective of the communication *link-layer*. This is in contrast to existing channel models, which characterize the wireless channel at the *physical-layer*. Specifically, we modeled the wireless link in terms of two 'effective capacity' functions; namely, the probability of non-empty buffer $\gamma^{(c)}(u)$ and the QoS exponent $\theta^{(c)}(u)$. The QoS exponent is the inverse of a function which we call *effective capacity* (EC). The EC link model is the dual of the effective bandwidth source traffic model, used in wired networks. Furthermore, we developed a simple and efficient algorithm to estimate the EC functions $\{\gamma^{(c)}(u), \theta^{(c)}(u)\}$. Simulation results show that the actual QoS metric is closely approximated, by the QoS metric predicted by the EC link model and its estimation algorithm, under various scenarios.

We have provided key insights about the relations between the EC link model and the physical-layer channel, *i.e.*, $\gamma^{(c)}(u)$ corresponds to the marginal CDF (*e.g.*, Rayleigh/Ricean distribution) while $\theta^{(c)}(u)$ is related to the Doppler spectrum. The EC link model has been justified not only from a theoretical viewpoint (*i.e.*, Markov property of fading channels) but also from an experimental viewpoint (*i.e.*, the delay-violation probability does decay exponentially with the delay).

The QoS metric considered can be easily translated into traffic envelope and service curve characterizations, which are popular in wired networks, such as ATM and IP, to provide guaranteed service. Therefore, we believe that the EC link model, which was specifically constructed keeping in mind this QoS metric, will find wide applicability in future wireless networks that need QoS provisioning.

In summary, our EC link model has the following features: simplicity of implementation, efficiency in admission control, and flexibility in allocating bandwidth and delay for connections. In addition, our link model provides a general framework, under which physical-layer fading channels such as AWGN, Rayleigh fading, and Ricean fading channels can be studied.

Armed with the new link model, we are now investigating its use in designing admission control, resource reservation, and scheduling algorithms, for efficient support of a variety of traffic flows that require guaranteed QoS. We are also exploring the implementation of our link model in 3G wireless systems, and its implications in QoS support for such networks.

Appendix

We show that the $\{\gamma^{(c)}(\mu), \theta^{(c)}(\mu)\}$ functions that specify our effective capacity link model, can be easily used to obtain the service curve specification $\Psi(t) = \{\sigma^{(c)}, \lambda_s^{(c)}\}$. The parameter $\sigma^{(c)}$ is simply equal to the source delay requirement D_{max}. Thus, only the channel sustainable rate $\lambda_s^{(c)}$ needs to be estimated. $\lambda_s^{(c)}$ is the source rate μ at which the required QoS (delay-violation probability ε) is achieved.

The following binary search procedure estimates $\lambda_s^{(c)}$ for a given (unknown) fading channel and source specification $\{D_{max}, \varepsilon\}$. In the algorithm, ϵ_e is the error between the target and the estimated $\sup_t Pr\{D(t) \geq D_{max}\}$, ϵ_t the precision tolerance, μ the source rate, μ_l a lower bound on the source rate, and μ_u an upper bound on the source rate.

Algorithm 1 (Estimation of the channel sustainable rate $\lambda_s^{(c)}$)

```
/* Initialization */
Initialize $\varepsilon$, $\epsilon_t$, and $\epsilon_e$;
/* E.g., $\varepsilon = 10^{-3}$; $\epsilon_t = 10^{-2}$; $\epsilon_e = 1$; */
$\mu_l := 0$; /* An obvious lower bound on the rate */
$\mu_u := r_{awgn}$; /* An obvious upper bound on the rate is the AWGN capacity */
$\mu := (\mu_l + \mu_u)/2$;
/* Binary search to find a $\lambda_s^{(c)}$ that is conservative and within $\epsilon_t$ */
While (($\epsilon_e/\varepsilon > \epsilon_t$) or ($\epsilon_e < 0$ )) {
    The data source transmits at the output rate $\mu$;
    Estimate $\hat{\gamma}$ and $\hat{\theta}$ using (19) to (22);
    Use Eq. (23) to obtain $\sup_t Pr\{D(t) \geq D_{max}\}$;
    $\epsilon_e := \varepsilon - \sup_t Pr\{D(t) \geq D_{max}\}$;
    if ($\epsilon_e \geq 0$) { /* Conservative */
        if ($\epsilon_e/\varepsilon > \epsilon_t$) { /* Too conservative */
            $\mu_l := \mu$; /* Increase rate */
            $\mu := (\mu_l + \mu_u)/2$;
        }
    }
    else { /* Optimistic */
        $\mu_u := \mu$; /* Reduce rate */
        $\mu := (\mu_l + \mu_u)/2$;
    }
}
$\lambda_s^{(c)} := \mu$.
```

Algorithm 1 uses a binary search to find the channel sustainable rate $\lambda_s^{(c)}$. An alternative approach is to use a parallel search, such as the one described in Ref. [11]. A parallel search would require more computations, but would converge faster.

References

[1] ATM Forum Technical Committee, "Traffic management specification (version 4.0)," ATM Forum, 1996.

[2] R. Braden (Ed.), L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) version 1, functional specification," *RFC 2205,* Internet Engineering Task Force, Sept. 1997.

[3] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[4] G. L. Choudhury, D. M. Lucantoni, W. Whitt, "Squeezing the most out of ATM," *IEEE Transactions on Communications,* vol. 44, no. 2, pp. 203–217, Feb. 1996.

[5] A. E. Eckberg, "Approximations for bursty and smoothed arrival delays based on generalized peakedness," in *Proc. 11th International Teletraffic Congress,* Kyoto, Japan, Sept. 1985.

[6] P. Ferguson and G. Huston, *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* Wiley, 1998.

[7] R. Guerin and V. Peris, "Quality-of-service in packet networks: basic mechanisms and directions," *Computer Networks and ISDN,* vol. 31, no. 3, pp. 169–179, Feb. 1999.

[8] S. Hanly and D. Tse, "Multi-access fading channels: part II: delay-limited capacities," *IEEE Trans. on Information Theory,* vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[9] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications,* Wiley, 2000.

[10] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Personal Communications Magazine,* pp. 4–9, Dec. 1996.

[11] B. L. Mark and G. Ramamurthy, "Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks," *IEEE/ACM Trans. on Networking,* vol. 6, no. 6, pp. 811–827, Dec. 1998.

[12] T. S. Rappaport, *Wireless Communications: Principles & Practice,* Prentice Hall, 1996.

[13] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," *RFC 2212,* Internet Engineering Task Force, Sept. 1997.

[14] Q. Zhang and S. A. Kassam, "Finite-state markov model for Rayleigh fading channels," *IEEE Trans. Commun.,* vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[15] Z.-L. Zhang, "End-to-end support for statistical quality-of-service guarantees in multimedia networks," *Ph.D. Dissertation,* Department of Computer Science, University of Massachusetts, Feb. 1997.

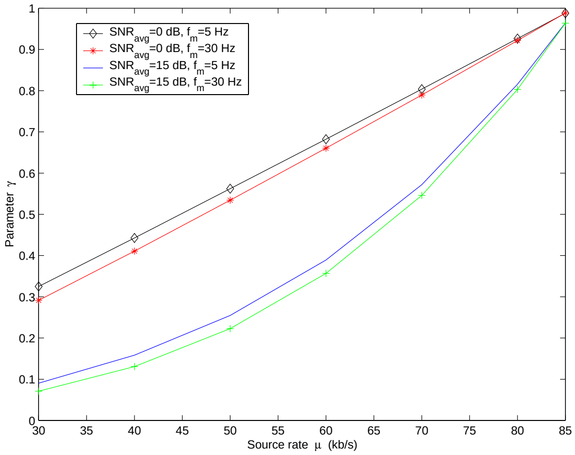


Figure 7: Estimated function $\hat{\gamma}(\mu)$ vs. source rate μ.

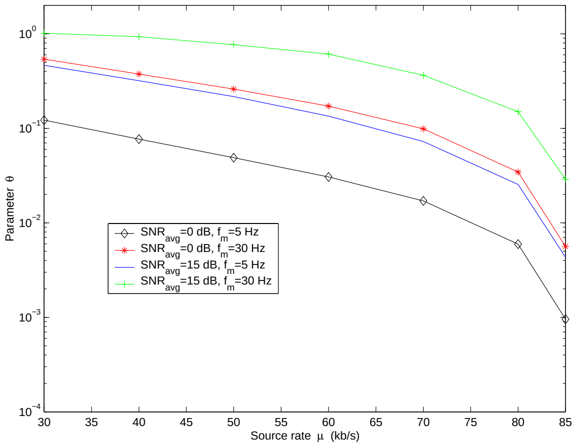


Figure 8: Estimated function $\hat{\theta}(\mu)$ vs. source rate μ.

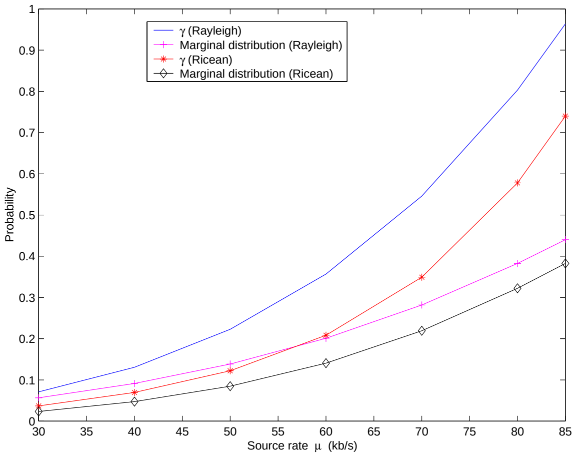


Figure 9: Marginal CDF and $\gamma(\mu)$ vs. source rate μ.

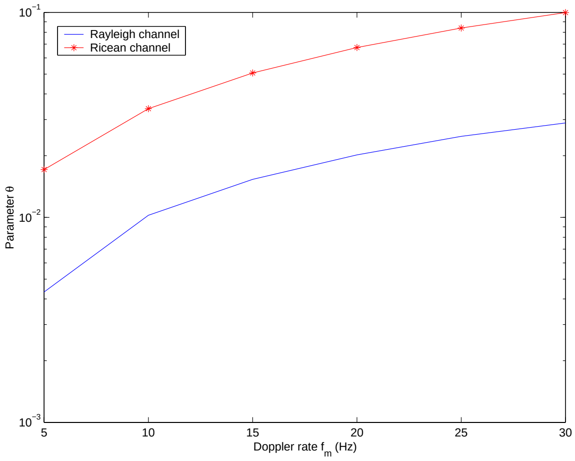
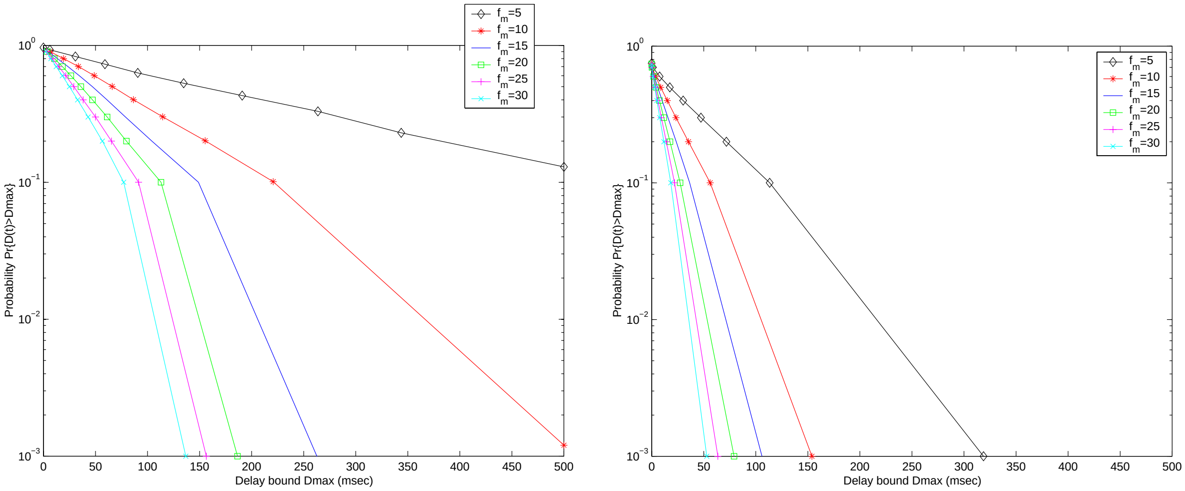


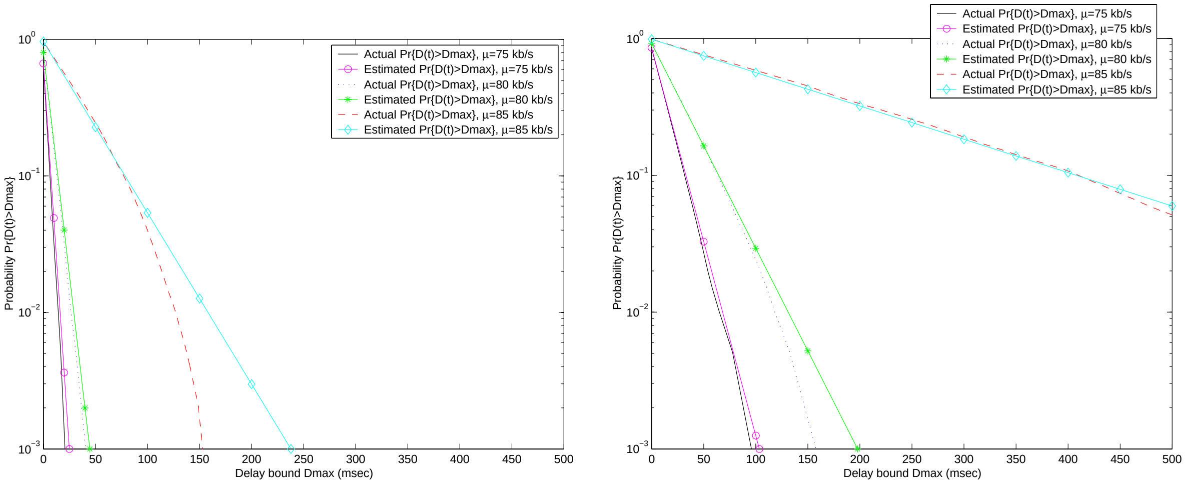
Figure 10: θ vs. Doppler rate f_m.



(a)

(b)

Figure 11: Actual delay-violation probability vs. D_{max}, for various Doppler rates: (a) Rayleigh fading and (b) Ricean fading.



(a) (b)

Figure 12: Prediction of delay-violation probability, when the average SNR is (a) 15 dB and (b) 0 dB.

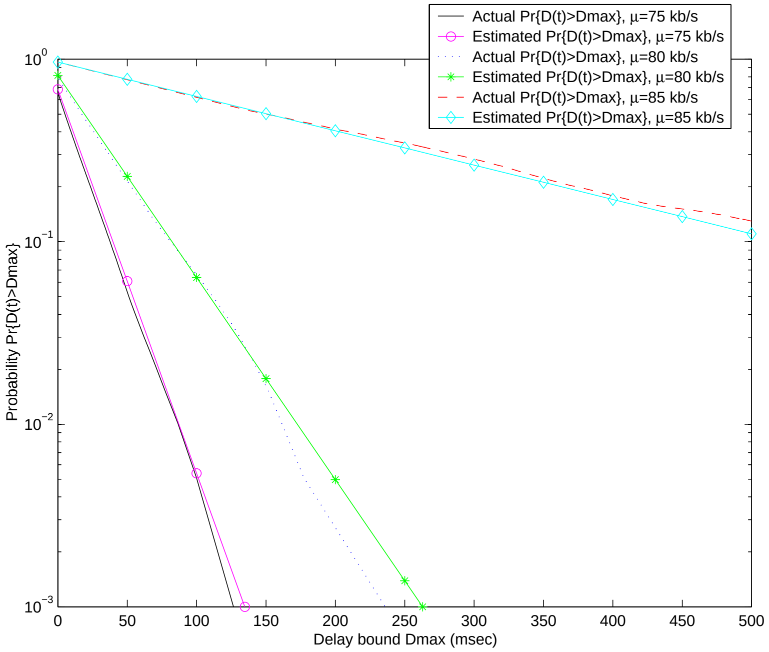


Figure 13: Prediction of delay-violation probability, when $f_m = 5$ Hz.