Fluid Analysis of Delay Performance for QoS Support in Wireless Networks*

Jeong Geun Kim and Marwan Krunz

Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85721
{jkkim, krunz}@ece.arizona.edu

Abstract

Providing quality of service (QoS) guarantees over wireless links requires thorough understanding and quantification of the interactions among the traffic source, the wireless channel, and the underlying error control mechanisms. In this paper, we account for such interactions in a network-layer model that we use to investigate the delay performance for an ON/OFF traffic stream transported over a wireless link. The capacity of this link fluctuates according to a fluid version of Gilbert-Elliot's model. We derive the packet delay distribution via two different approaches: uniformization and Laplace transform. Conputational aspects of both approaches are discussed. The delay distribution is then used to quantify the wireless effective bandwidth under a given delay guarantee. Numerical results and simulations are used to verify the adequacy of our analysis and to study the impact of error control and bandwidth allocation on the packet delay performance.

keywords: *Wireless networks, QoS, delay distribution, fluid analysis.*

1. Introduction

Recent trends in wireless networks indicate a desire to provide a flexible broadband wireless infrastructure that can support emerging multimedia services as well as traditional data services [1, 13]. In such a multi-service wireless environment, quality-of-service (QoS) guarantees are critical for real-time voice and video. In contrast to its wireline counterpart, the provisioning of QoS guarantees over wireless links is a more challenging problem whose difficulty stems from the need to explicitly consider the harsh radio-channel transmission characteristics and the underlying link-layer error control mechanisms. This difficulty is further compounded by host mobility and how it impacts the available bandwidth capacity. These issues indicate a clear need for a general QoS framework in the wireless environment.

QoS guarantees in wireless networks can be provided through a coordination between connection-level bandwidth reservation and packet-level scheduling. Most previous research on QoS over wireless networks has mainly focused on these issues. Levine et al. proposed the shadow cluster concept to estimate the bandwidth requirements of a wireless connection [9]. They indicated that this concept can be used in connection admission control to provide a specified call dropping probability. Reininger et al. identified the high variability of traffic dynamics of mobile multimedia applications as a function of time and space, and proposed a soft QoS control which allows bandwidth renegotiation according to the varying traffic conditions [14]. Capone and Stavrakakis investigated the region of supportable QoS vectors expressed in terms of packet dropping probability [3]. Their work provided insight into the resource management aspects for handling diverse QoS constraints, although the study was limited to unbuffered services. Lu et al. proposed a fair scheduling algorithm with adaptation to wireless networks that take into account bursty and location-dependent channel errors [11]. Although their work identified many practical issues, it did not address the interaction between packet scheduling and error control. In [7], the authors studied the concept of wireless effective bandwidth for a guaranteed packet loss rate.

The general goals of the underlying work are to study the delay performance over a wireless link and investigate its implication on optimal bandwidth allocation under delay guarantees. Our investigations are carried out for a single stream that is transported over a time-varying wireless link. If the link is used to transport more than one connection, then each connection is guaranteed a constant service rate during its active period (i.e., TDMA style). The outcome

*This research was supported by the National Science Foundation under CAREER Grant ANI-9733143.

of a packet transmission is determined by the state of the wireless channel and the error control schemes. This scenario encompasses point-to-point connections between mobile terminals (MT) and a base station (BS) in cellular communication systems.

To achieve our goals, we follow a fluid-based approach whereby the traffic source is modeled by an on-off fluid process and the channel is modeled by a fluid variant of Gilbert-Elliott's model. Using fluid-flow analysis, we compute the delay distribution for a single stream as a function of the traffic source, the service rate, the wireless channel, and the error control schemes. To obtain this distribution, we first evaluate the queue length distribution taking into account the channel behavior and the underlying error control schemes. Then, we provide two alternative approaches for obtaining the delay distribution via the uniformization and Laplace transform techniques. In the case of the uniformization approach, we are able to derive a closed-form expression for the delay distribution. The computational aspects of both approaches are compared. Our analytical results are used to obtain the wireless effective bandwidth *under a given delay constraint*. Note that the notion of effective bandwidth has been traditionally investigated in wireline [4] and wireless [7] networks for a given packet loss rate. We also study the optimal error control strategy that minimizes the effective bandwidth while guaranteeing a given delay requirement (expressed as a percentile). Extensive simulations are conducted to verify the goodness of our analytical results.

The rest of the paper is organized as follows. In Section 2, we describe the wireless link model. Analysis of the delay performance is provided in Section 3. Numerical results and simulations are reported in Section 4, followed by concluding remarks in Section 5.

2. Wireless Link Model

2.1. Framework

In order to analyze the packet-level performance of a wireless link, we consider the framework shown in Figure 1. This framework was used earlier to study the packet loss performance and the corresponding effective bandwidth under the same problem setting [7]. In this framework, traffic streams from one or more connections are fed into a finite-size FIFO buffer. A constant service rate c (in packets/second) is assigned to the wireless connection, but the actual drain rate observed at the buffer is reduced due to retransmissions and FEC overhead. The actual service rate will be discussed in the following section. In our study, we consider hybrid ARQ/FEC error control in which the cyclic redundancy check (CRC) code is applied first to a packet, followed by FEC. We assume that the CRC code can alone detect almost all bit errors in a packet. In contrast, only a subset of the errors can be corrected by FEC. In addition, we impose a limit on the number of packet transmissions. Imposing such a limit can be used to provide delay guarantees for real-time traffic. Once a packet hits the limit, it will be discarded. For simplicity, we ignore the overhead of the medium access control (MAC) layer.

The above model has three control parameters: the service rate (or assigned bandwidth), the FEC code rate, and the limit on the number of transmissions. These parameters can be adjusted during connection setup to satisfy certain QoS requirements. ¿From the network point of view, the selection of these parameters is very crucial and requires thorough understanding of their impact on the packet-level performance. The main theme of this study is to investigate the packet-level performance of a wireless link as a function of the assigned bandwidth, the limit on transmissions, and error control schemes.

2.2. Queueing Model

In this section, we describe the queueing model that is used to analyze the packet delay over a wireless link. The source is characterized by an on-off fluid process with peak rate r. Its on and off periods are exponentially distributed with means $1/\alpha$ and $1/\beta$, respectively. The wireless channel is modeled using a fluid version of Gilbert-Elliott (GE) model which is often used to investigate the performance over wireless links [6]. As explained in Figure 2, the GE model is Markovian with two alternating states: *Good* and *Bad*. The bit error rates (BER) during the Good and Bad states are given by P_{eg} and P_{eb}, respectively, where $P_{eg} \ll P_{eb}$. The durations of the Good and Bad states are exponentially distributed with means $1/\delta$ and $1/\gamma$, respectively.

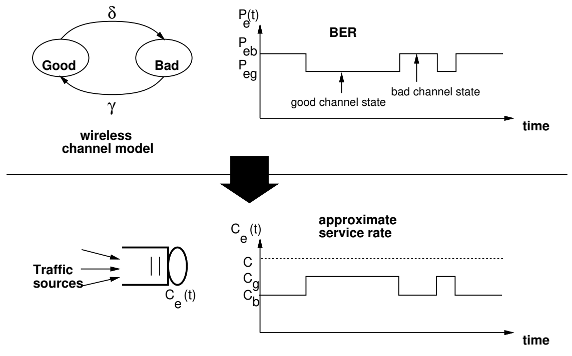

Figure 2. Wireless channel model and corresponding service rate model.

The FEC capability in the underlying hybrid ARQ/FEC mechanism is characterized by three parameters: the number of bits in a code block (n), the number of payload bits

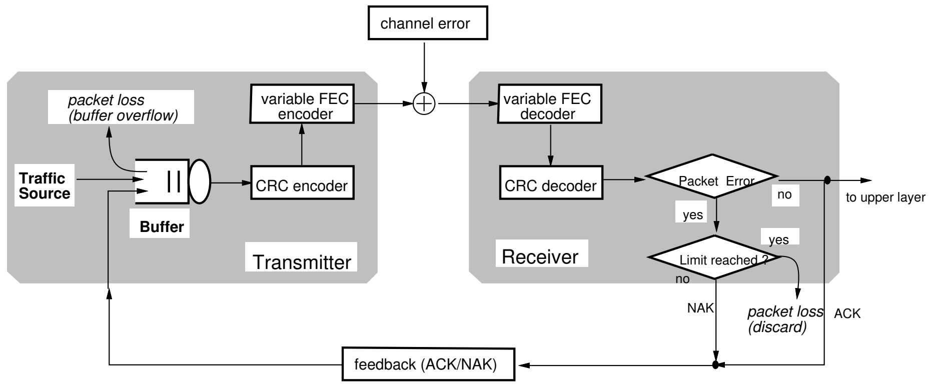

Figure 1. Framework for analyzing the performance over a wireless link.

(k), and the maximum number of correctable bits in a code block (τ). Note that n consists of the k payload bits and the extra parity bits. The FEC code rate $e(\tau)$ is defined as

$$e(\tau) = \frac{k}{n(\tau)}.$$

Assuming that a FEC code can correct up to τ bits and that bit errors during a given channel state are independent, the probability that a packet contains a non-correctable error is given by:

$$P_c(p_b, \tau) = \sum_{j=\tau+1}^{n(\tau)} \binom{n(\tau)}{j} p_b^j (1 - p_b)^{n(\tau)-j} \tag{1}$$

where p_b is the bit error probability; $p_b \in \{P_{eg}, P_{eb}\}$. To account for the FEC overhead, we obtain the actual service rate c_e observed at the output of the buffer:

$$c_e = c \cdot e(\tau) \tag{2}$$

where c is the bandwidth assigned to the connection.

The *exact* behavior of ARQ and FEC in the underlying queueing model is difficult to analyze. To obtain analytically tractable results, we assume that the packet departure process follows a fluid process with a service rate that is modulated by the channel state (see Fig. 2). This approximation implies that there are two deterministic service rates: c_g during Good states and c_b during Bad states. We assume that the feedback delay for sending an acknowledgment from a given receiver to the sender is smaller than the minimum time between two successive transmissions to that receiver. This assumption is reasonable in a TDMA environment, where the channel capacity is being shared by several connections (destinations, MTs). Each connection is assigned one or more slots within a TDMA frame. Slot assignment reflects the constant service rate that is allocated to a connection[1]. A packet is successively retransmitted until it is correctly received at the destination or until the limit on the number of retransmissions is reached. In this scenario, the total time needed to successfully deliver a packet conditioned on the channel state follows a truncated geometric distribution. Let N_{tr} denote the number of retransmissions (including the first transmission) until a packet is successfully received or is discarded because it reached the limit on retransmissions. For a given packet error probability P_c and a limit on transmissions N_l, the expected value of N_{tr} is given by:

$$E[N_{tr}] = \frac{1 - P_c^{N_l}}{1 - P_c}. \tag{3}$$

Thus, c_g and c_b correspond to the mean transmission rates of the truncated geometric trials with parameters $(P_{c,g}, N_l)$ and $(P_{c,b}, N_l)$, respectively, where $P_{c,g}$ and $P_{c,b}$ are the packet error probabilities in Good and Bad states, respectively, given by (1). Formally,

$$c_g = \frac{c \cdot e(\tau) \cdot (1 - P_{c,g})}{1 - P_{c,g}^{N_l}} \tag{4}$$

$$c_b = \frac{c \cdot e(\tau) \cdot (1 - P_{c,b})}{1 - P_{c,b}^{N_l}}. \tag{5}$$

where $P_{c,g} = P_c(P_{e,g}, \tau)$ and $P_{c,b} = P_c(P_{e,b}, \tau)$.

3. Analysis of Delay Performance

3.1. Queue Length Distribution

Following the discussion in the previous section, we construct a Markovian queueing system with four states as shown in Figure 3. Let S denote the state space. Thus,

$$S = \{(0, g), (0, b), (1, g), (1, b)\} \tag{6}$$

where 0 and 1 denote the on and off states of a traffic source, respectively, and g and b denote Good and Bad channel states, respectively.

[1] Guaranteeing a constant service rate to a connection can be achieved by periodic assignment of slots in a TDMA frame.

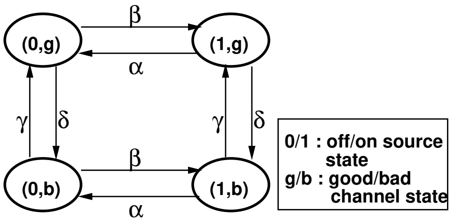

Figure 3. State transition diagram.

Following a standard fluid approach (see [2], for example), the evolution of the buffer content can be described by the following differential equation:

$$\frac{d\mathbf{\Pi}(x)}{dx}\boldsymbol{D} = \mathbf{\Pi}(x)\boldsymbol{M} \tag{7}$$

where $\boldsymbol{D} \triangleq \text{diag}[-c_g, -c_b, r - c_g, r - c_b]$, $\mathbf{\Pi}(x) \triangleq [\ \Pi_{0,g}(x) \quad \Pi_{0,b}(x) \quad \Pi_{1,g}(x) \quad \Pi_{1,b}(x)\ ]$,
$\Pi_s(x) \triangleq \Pr\{\text{buffer content} \leq x \text{ and the system } s \in S\}$, and $\boldsymbol{M}$ is the generator matrix of the underlying Markov chain:

$$\boldsymbol{M} = \begin{bmatrix} -(\beta+\delta) & \delta & \beta & 0 \\ \gamma & -(\beta+\gamma) & 0 & \beta \\ \alpha & 0 & -(\alpha+\delta) & \delta \\ 0 & \alpha & \gamma & -(\alpha+\gamma) \end{bmatrix}.$$

Throughout the paper, matrices and vectors are boldfaced.

The solution of (7) corresponds to the solution of the eigenvalue/eigenvector problem:

$$z\boldsymbol{\phi}\boldsymbol{D} = \boldsymbol{\phi}\boldsymbol{M} \tag{8}$$

which is generally given by

$$\mathbf{\Pi}(x) = \sum_{z_i \leq 0} a_i \exp(z_i x)\phi_i \tag{9}$$

where a_i's are constant coefficients and the pairs $(z_i, \boldsymbol{\phi}_i)$, $i = 1, 2, \cdots$, are the eigenvalues and the right eigenvectors of the matrix $\boldsymbol{M}\boldsymbol{D}^{-1}$ [2, 12]. In order to solve (8), we follow the approach used in [12]. The details of how $\mathbf{\Pi}(x)$ is computed are given in [7]. In the following section, $\mathbf{\Pi}(x)$ is used to obtain the delay distribution.

3.2. Delay Distribution

In fluid queueing models with an error-free channel and constant service rate, e.g., ATM link, the delay distribution can be directly obtained from the queue length distribution [5]. However, the scenario we consider in this study includes a time-varying wireless channel that is being approximated by a two-state Markov modulated fluid process. In this case, the packet delay distribution is much more difficult to obtain since one has to take into account the time-varying service rate as well as the queue length.

We assume an infinite-capacity buffer. Let D denote the delay experienced by an arriving packet. Let $C(t)$ denote the accumulative amount of service during a period of length t:

$$C(t) \triangleq \int_0^t c(s)ds$$

where $c(s)$ is the service rate at time s. The channel state at time t is denoted by $h(t) \in \{g, b\}$, where g and b denote Good and Bad states, respectively. The probability that the delay seen by a fluid atom is less than or equal to t is equal to the probability that $C(t)$ is greater than or equal to the queue length at the instant of the packet arrival Q_0. Thus, we have

$$\begin{aligned} \Pr[D \leq t] &= \Pr[C(t) \geq Q_0] \\ &= \sum_{i \in S} \int_{0-}^{\infty} \Pr[C(t) \geq x | i, Q_0 = x]\pi_i(x)dx \\ &= \frac{r}{T}\int_{0-}^{\infty} \Pr[C_g(t) \geq x]\pi_{1,g}(x) \\ &\qquad + \Pr[C_b(t) \geq x]\pi_{1,b}(x)dx \end{aligned} \tag{10}$$

where T is the throughput, π_i is the pdf of the queue length in a state i, $i \in S$, and

$$C_i(t) \triangleq \int_0^t c(s)ds \quad \text{given } h(0) = i, \text{ for } i \in \{g, b\}.$$

The quantity $r\pi(x)/T$ represents the fraction of carried flow that arrives at the queue when its content is x. For an infinite-capacity buffer, the throughput T is given by:

$$T = r(w_{1,g} + w_{1,b}). \tag{11}$$

In order to obtain $\Pr[C_i(t) \leq x]$, $i \in \{g, b\}$, we provide two methods: direct calculation using Laplace transform and uniformization. The equivalence of these approaches will be verified using numerical examples.

Laplace Transform Approach

For convenience, we transform the random variable $C_i(t)$ to $\bar{C}_i(t)$ defined by:

$$\bar{C}_i(t) \triangleq C_i(t) - c_b t, \quad i \in \{g, b\}.$$

By this transformation, $\bar{C}_i(t)$ is the accumulative service resulting from a "normalized" channel with service rates $c_g - c_b$ (during Good states) and 0 (during Bad states). Note that the minimum amount of accumulative service in a period of length t is $c_b t$. Thus,

$$\Pr[C_i(t) \geq x] = \begin{cases} 1, & \text{if } x < c_b t \\ 1 - \Pr[\bar{C}_i(t) \leq x - c_b t], & \text{if } x \geq c_b t. \end{cases}$$

The following proposition gives the probabilities $\Pr[C_g(t) \geq x]$ and $\Pr[C_b(t) \geq x]$ by solving the partial differential equations (PDE) for the consumption rate $C(t)$.

Proposition 3.1 *The probabilities* $\Pr[C_i(t) \geq x], i \in \{g,b\}$ *when* $x \geq c_b t$*, are given by*

$$\Pr[C_g(t) \geq x] = e^{-\delta\hat{x}} \left(e^{-\gamma(t-\hat{x})} J_0(2\sqrt{-\delta\gamma\hat{x}(t-\hat{x})}) + \sum_{n=0}^{\infty} \frac{\delta\hat{x}}{(n!)^2} \Gamma(n+1, \gamma(t-\hat{x})) \right) \quad (12)$$

$$\Pr[C_b(t) \geq x] = e^{-\delta\hat{x}} \sum_{n=0}^{\infty} \frac{\delta\hat{x}}{(n!)^2} \Gamma(n+1, \gamma(t-\hat{x})) \quad (13)$$

where $\hat{x} = (x - c_b t)/(c_g - c_b)$*,* $J_0(z)$ *is the Bessel function of order zero given by*

$$J_0(z) = \sum_{n=0}^{\infty} \frac{(-1)^n (z/2)^{2n}}{(n!)^2},$$

and $\Gamma(n,z)$ *is the incomplete gamma function given by:*

$$\Gamma(n,z) = \int_0^z e^{-x} x^{n-1} dx.$$

Proof. See [8].

Equation (13) in Proposition 3.1 is substituted into (10) to evaluate the delay distribution. Some numerical complexity is associated with the infinite sums in Equation (12) and (13). We observed that the values of these infinite sums converge fast for moderate values of n, e.g., $n = 20$.

Uniformization Approach

As a second approach to obtaining $\Pr[C_i(t) \leq x], i \in \{g,b\}$, we use the uniformization approach. In continuous-time Markov chains, uniformization is a technique for uniformizing the transition rates between states by introducing transitions from a state to itself [15, 16].

Let t_g and t_b denote the accumulative sojourn times of Good and Bad channel states during an interval of length t, respectively. That is,

$$t_g \triangleq \int_0^t \mathbf{1}_{\{h(s)=g\}} ds$$
$$t_b \triangleq \int_0^t \mathbf{1}_{\{h(s)=b\}} ds.$$

Then, the accumulative service $C(t)$ is given by

$$C(t) = c_g t_g + c_b t_b, \quad 0 \leq t_g, t_b \leq t. \quad (14)$$

Since $t = t_g + t_b$, $C(t)$ can be expressed as

$$C(t) = c_g t_g + c_b(t - t_g) \quad \text{or} \quad C(t) = c_g(t - t_b) + c_b t_b. \quad (15)$$

In [8], we provide the probability distribution of t_g and t_b conditioned on the channel state. Then, the probability $\Pr[C_i(t) \geq x], i \in \{g,b\}$ can be directly obtained from (15).

Proposition 3.2 *The probabilities* $\Pr[C_g(t) \geq x]$ *and* $\Pr[C_b(t) \geq x]$ *are given by*

$$\Pr[C_g(t) \geq x] = 1 - e^{-(\delta+\gamma)t} \sum_{n=1}^{\infty} \frac{(\delta t)^n}{n!} \sum_{k=1}^{n} \binom{n}{k-1} \cdot \left(\frac{\gamma}{\delta}\right)^{k-1} \sum_{i=k}^{n} \binom{n}{i} \chi^i (1-\chi)^{n-i} \quad (16)$$

and $$\Pr[C_b(t) \geq x] = e^{-(\delta+\gamma)t} \sum_{n=1}^{\infty} \frac{(\gamma t)^n}{n!} \sum_{k=1}^{n} \binom{n}{k-1} \cdot \left(\frac{\delta}{\gamma}\right)^{k-1} \sum_{i=k}^{n} \binom{n}{i} \chi^{n-i} (1-\chi)^i \quad (17)$$

where $\chi = \frac{x - c_b t}{(c_g - c_b)t}$.

Proof. See [8].

The equations in Proposition 3.2 also have some numerical issues due to the presence of multiple sums. The triple sums in the equations cause significant complexity. However, observing the duplicate computation in the last sum for consecutive indexes k's, we can achieve a significant reduction in the computation time.

Up to this point, we have discussed the alternative approaches to obtaining the probability $\Pr[C_i(t) \geq x], i \in \{g,b\}$. In the following, the results from Proposition 3.2 are used to obtain the delay distribution. The following proposition gives a closed-form expression for delay distribution by substituting (16) and (17) into (10).

Proposition 3.3

$$\begin{aligned}
\Pr[D \leq t] = &\frac{r}{T}(\Pi_{1,g}(c_g t) + \Pi_{1,b}(c_b t)) \\
&- \frac{r}{T} e^{-(\delta+\gamma)t} \sum_{n=1}^{\infty} \frac{(\delta t)^n}{(n+1)!} \sum_{k=1}^{n} \binom{n}{k-1} \left(\frac{\gamma}{\delta}\right)^{k-1} \\
&\cdot \sum_{i=k}^{n} (c_g - c_b)t \sum_{l} a_l e^{z_l c_b t} \Phi(i+1; n+2; z_l(c_g - c_b)t) \\
&+ \frac{r}{T} e^{-(\delta+\gamma)t} \sum_{n=1}^{\infty} \frac{(\gamma t)^n}{(n+1)!} \sum_{k=1}^{n} \binom{n}{k-1} \left(\frac{\delta}{\gamma}\right)^{k-1} \\
&\cdot \sum_{i=k}^{n} (c_g - c_b)t \sum_{m} a_m e^{z_m c_b t} \\
&\cdot \Phi(n-i+1; n+2; z_m(c_g - c_b)t)
\end{aligned} \quad (18)$$

where $\pi_{1,g}(x) = \sum_l a_l e^{z_l x}$, $\pi_{1,b}(x) = \sum_m a_m e^{z_m x}$, *the indexes* l, m *are used to index the negative eigenvalues, and*

$$\Phi(x; y; z) \quad \stackrel{\triangle}{=} \quad \sum_{k=0}^{\infty} \frac{(x)_k}{(y)_k} \frac{z^k}{k!}$$

with $(a)_n \stackrel{\triangle}{=} a(a+1)\cdots(a+n-1)$.

Proof. See [8].

Since Equation (18) has a similar numerical structure to the expressions in Proposition 3.2, we can reduce the computation time by avoiding the duplicate sums as mentioned previously. We also obtained the delay distribution using Proposition 3.1. However, the resulting expression contains multiple sums, and hence provides no advantages over the uniformization approach. Accordingly, to obtain the delay distribution in the first approach (Proposition 3.1), we rely on numerical integration. This is mainly used in contrasting our two analysis approaches; most numerical results in Section 4 are based on Proposition 3.3.

Wireless Effective Bandwidth

The notion of effective bandwidth has been traditionally employed to provide a guaranteed packet loss rate in wireline [4] and wireless [7]. In this study, we extend this notion for a wireless connection under probabilistic delay constraints.

We defined the *wireless effective bandwidth* c_{eb} under the delay constraint $\Pr[\text{delay} > t] = \varepsilon$ as follows:

$$c_{eb} \stackrel{\triangle}{=} \min\{c | c \text{ satisfies } \Pr[\text{delay} > t] = \varepsilon\} \qquad (19)$$

where c is the service rate. In contrast to wireline effective bandwidth, the wireless effective bandwidth is configured along with the optimal number of correctable bits which minimizes the use of bandwidth while providing the requested reliability at the physical link. In Section 4, we provide some numerical examples related to this concept and investigate the characteristics of the pair of QoS parameters (c_{eb}, τ) in more detail.

4. Numerical Results and Discussion

In this section, we present numerical examples based on our analytical results. We verify the adequacy of these results by contrasting them against more realistic simulations.

Similar to the analysis, the simulation results are obtained using on-off traffic sources with exponentially distributed on and off periods. The ARQ retransmission process is simulated in a more realistic manner, whereby a packet is transmitted repeatedly until it is received with no errors or until it reaches the limit on the number of transmissions. The probability of a packet error is computed from (1) for both channel states. Transitions between Good and Bad states are assumed to occur only at the beginning of a packet transmission slot. A packet is retransmitted if it has uncorrectable errors. It is assumed that the propagation delay is small, so that the ACK/NAK message for a packet is received at the sender before the next attempt of transmission. Finally, we use an infinite-capacity buffer in our simulations.

In our experiments, we vary the BER during the Bad state (P_{eb}) and fix the BER during the Good state at $P_{eg} = 10^{-6}$. We set the mean of the off period to ten times that of the on period. In addition, we take the parameters related to the wireless channel from [6]. We adopt Bose-Chaudhuri-Hocquenghem (BCH) code [10] for FEC . We consider fixed packet sizes, e.g., ATM cells. Since we treat the CRC code as part of the payload, the FEC code is applied to 424-bit blocks (i.e., $k = 424$ bits). In [8], a table is available to show the size, the code rate, and the number of correctable bits of the BCH code used in our examples. All simulations are reported with 95% confidence intervals. For the delay distribution, 10^7 to 4×10^7 samples were needed in the simulations. Table 1 summarizes the values of the various parameters in the simulations and numerical examples. For the parameters c, P_{eb}, τ, and N_l, the values in the parenthesis are assumed unless specified otherwise.

Figure 4 depicts the complementary delay distribution. We vary the service rate (c) from 800 to 1200 packets/sec while fixing the other parameters at $P_{eb} = 10^{-2}, \tau = 7$, and $N_l = \infty$. The difference between the analytical and simulation results is quite negligible for all service rates. We observe a slight deviation at the tail part of the distribution. However, it is associated with the number of samples taken from the simulation. For $c = 1200$, we generated 4×10^7 packets to obtain the shown results. Note that the simulation is based on the realistic scenario in which the packet is transmitted until it is successfully transmitted or until it reaches the limit on the retransmission, whereas the analytical results are based on the fluid approximation.

Figure 5 shows the effective bandwidth as a function of the number of correctable bits (τ) for three target delay constraints $\Pr[\text{delay} > \varepsilon]$, with $\varepsilon = 0.01, 0.05, 0.1$. Expectedly, more bandwidth is needed to achieve a more stringent delay guarantee. Interestingly, we observe that a very large amount of bandwidth is required when only ARQ ($\tau = 0$) is used for error control. Thus, the use of FEC is essential to achieve efficient bandwidth allocation with delay guarantees. The figure clearly indicates that there is an optimal τ ($\tau = 7$ in this example) for a given BER that satisfies a delay QoS constraint while minimizing the use of bandwidth.

The optimal number of correctable bits as a function of the BER of the Bad state is shown in Fig. 6. The target delay constraint is fixed at $\Pr[\text{delay} > 0.01] = 0.25$. We vary the $P_{e,b}$ from 0.001 to 0.0158. For each BER, we observe the

Parameter	Symbol	Value
source peak rate	r	1 Mbps (or 2604.1667 packets/sec)
service rate	c	$100 - 8000$ packets/sec (1000)
mean on period	$1/\alpha$	0.02304 sec
mean off period	$1/\beta$	0.2304 sec
mean Good channel period	$1/\delta$	0.1 sec
mean Bad channel period	$1/\gamma$	0.0333 sec
BER in Good channel state	P_{eg}	10^{-6}
BER in Bad channel state	P_{eb}	$10^{-2} - 10^{-5}$ (10^{-2})
number of correctable bits	τ	$0 - 20$ (7)
limit on transmissions	N_l	$1 - \infty$ (∞)

Table 1. Parameter values used in the simulations and numerical results.

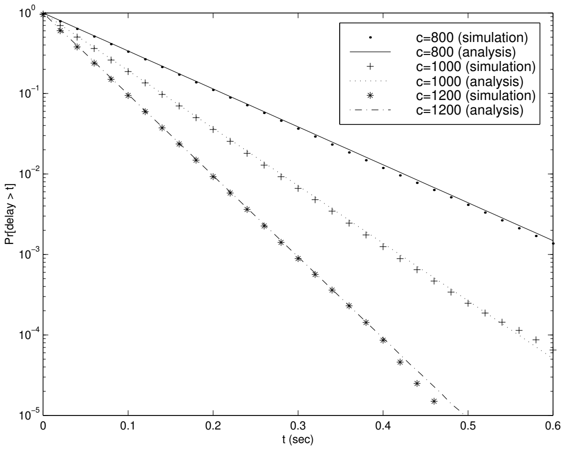

Figure 4. Complementary delay distribution for different service rates.

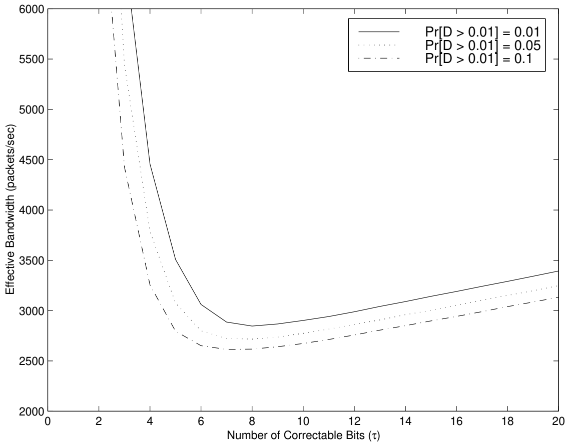

Figure 5. Effective bandwidth versus τ for 1 target delay constraint $\Pr[\text{delay} > 0.01] = \varepsilon$ $(\varepsilon = 0.01, 0.05, 0.1)$.

optimal number of correctable bits.

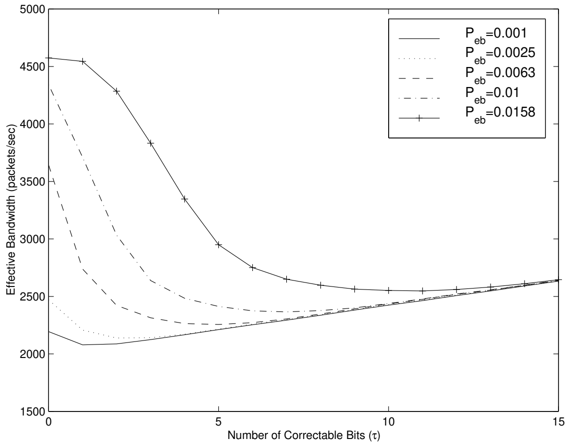

Figure 6. Effective bandwidth versus τ for different BER's.

Figure 7 shows the effective bandwidth versus the target delay constraints $\varepsilon = \Pr[\text{delay} > t]$, for $t = 0.001, 0.005, 0.01, 0.1$ (sec). As expected, more bandwidth is required for a more stringent delay requirement, i.e., smaller $\Pr[\text{delay} > t]$ at a fixed t. Notice that the minimum and maximum values of the effective bandwidth are $c = 282$ and $c = 3250$ (packets/sec), respectively. The minimum value corresponds to the minimum service rate satisfying the stability condition of the queue, whereas the maximum value is the minimum service rate satisfying the condition $c_b > r$, i.e., $\min\{c | c_b > r\}$. Considering the the source peak rate $r = 2604.1667$ (packets/sec), the maximum effective bandwidth is surprisingly greater than the source peak rate. This is in disagreement with the well-known fact that the wireline effective bandwidth is bounded by the source peak rate. In the underlying wireless scenario, the assigned service rate is reduced due to the packet errors and FEC overhead, and thus the principle in wireline effective bandwidth cannot be applied here. Another interesting observation in this figure is associated with the shape of the curves with decreasing $\Pr[\text{delay} > t]$. Consider the case of $t = 0.001$. To improve the QoS from $\Pr[\text{delay} > 0.001] = 1.0$ to $\Pr[\text{delay} > 0.001] = 0.8$, we need to assign an extra bandwidth 1950.0 (packets/sec), whereas we need 143.0 (packets/sec) for $t = 0.1$. That is, the behavior of the effective bandwidth depends on the selection of t as well as the constraint $\Pr[\text{delay} > t]$.

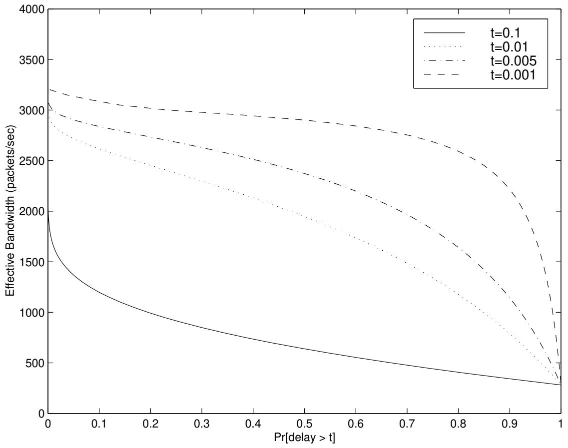

Figure 7. Effective bandwidth versus target delay constraints $\Pr[\text{delay} > t]$.

5. Conclusions

In this paper, we investigated the delay performance for an on/off source transported over a wireless channel. Simple yet accurate fluid models were used to capture the bursty nature of the arriving traffic and the channel's time-varying error characteristics. Error control schemes (ARQ and FEC), which are essential elements of any wireless packet network, were incorporated. We obtained the delay distribution using two alternative approaches: Laplace transform and uniformization. The solution was then used to obtain the wireless effective bandwidth (defined here as the minimum amount of bandwidth required to satisfy a given probabilistic delay constraint), which can be used as a valuable tool in resource allocation and admission control in wireless networks. Our analytical results were validated by contrasting them with simulations. It was observed that the analytically obtained delay distribution is quite accurate over a wide range of parameters. In a future work, we plan to investigate the system capacity of the wireless link, which is constrained by diverse QoS parameters, e.g., delay, loss, and jitter, of multiple connections over a shared channel.

References

[1] A. Acampora. Wireless ATM: a perspective on issues and prospects. *IEEE Pers. Commun.*, 3(4):8–17, Aug. 1996.

[2] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data–handling system with multiple sources. *Bell Syst. Tech. J.*, 61:1871–1894, 1982.

[3] J. Capone and I. Stavrakakis. Achievable QoS and scheduling policies in integrated services wireless networks. *Perform. Eval.*, 27/28(1):347–365, Oct. 1996.

[4] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, 1(3):329–343, June 1993.

[5] A. I. Elwalid and D. Mitra. Statistical multiplexing with loss priorities in rate-based congestion control of high-speed networks. *IEEE Trans. Commun.*, 42(11):2989–3002, Nov. 1994.

[6] N. Guo and S. D. Morgera. Frequency-hopped ARQ for wireless network data services. *IEEE J. Select. Areas Commun.*, 12(8):1324–1336, Sept. 1994.

[7] J. G. Kim and M. Krunz. Effective bandwidth in wireless ATM networks. In *MobiCom '98*, pages 233–241, Oct. 1998.

[8] J. G. Kim and M. Krunz. Fluid analysis of delay and packet discard performance for QoS support in wireless networks. Technical Report CENG-TR-99-119, Department of ECE, University of Arizona, http://www.ece.arizona.edu/∼bnlab, Aug. 1999.

[9] D. A. Levine, I. F. Akyildiz, and M. Naghshineh. A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Trans. Networking*, 5(1):1–12, Feb. 1997.

[10] S. Lin and J. D. J. Costello. *Error Control Coding: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1983.

[11] S. Lu, V. Bharghavan, and R. Srikant. Fair scheduling in wireless packet networks. In *SIGCOMM '97*, Sept. 1997.

[12] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Prob.*, 20:646–676, 1988.

[13] D. Raychaudhuri and N. D. Wilson. ATM-based transport architecture for multiservices wireless personal communication networks. *IEEE J. Select. Areas Commun.*, 12(8):1401–1414, Oct. 1994.

[14] D. Reininger, R. Izmailov, B. Rajagopalan, M. Ott, and D. Raychaudhuri. Soft QoS control in the WATMnet broadband wireless system. *IEEE Pers. Commun.*, 6(1):34–43, Feb. 1999.

[15] S. Ross. *Stochastic Processes*. John Wiley & Sons, second edition, 1996.

[16] B. Sericola. Transient analysis of stochastic fluid models. *Performance Evaluation*, 32:245–263, 1998.