Providing Statistical Delay Guarantees in Wireless Networks

Shengquan Wang, Ripal Nathuji, Riccardo Bettati and Wei Zhao

Abstract

This work studies the delay performance of policed traffic to provide real-time guarantees over wireless networks. A number of models have been presented in the literature to describe wireless (radio or optical) networks in terms of the wireless channel and the underlying error control mechanisms. In this paper, we describe a general framework to incorporate such models into delay guarantee computations for real-time traffic. Static-priority scheduling is considered, and two different admission control mechanisms are used to achieve the trade-off between resource utilization and admission overhead.

1. Introduction

The convenience of wireless communications has led to an increasing use of wireless networks for both civilian and mission critical applications. Many of these kinds of applications require delay-guaranteed communications. In the following, we describe approaches to provide delay-guaranteed services in wireless networks.

A significant amount of work has been done on real-time communication over wired networks [3, 10, 12, 13, 19, 20]. Wireless networks, however, are substantially different from their wired counterparts, and technologies developed for wired networks cannot be directly adopted: In most wired network models for real-time systems, the communication links are assumed to have a fixed capacity over time. This assumption may be invalid in wireless (radio or optical) environments, where link capacities can be temporarily degraded due to fading, attenuation, and path blockage [15, 18]. In order to improve the performance of wireless links, error control schemes are used. Common error control methods used in wireless communications include forward error correction (FEC), automatic repeat request (ARQ) and their hybrids [2, 5].

The difficulty of provisioning real-time guarantees in wireless networks stems from the need to explicitly consider both the channel transmission characteristics and the underlying error control mechanisms. There is a large volume of literature dealing with the representation and analysis of channel models, and most of these models directly characterize the fluctuations of signals and provide an estimate of the performance characteristics such as symbol error rate vs. signal-to-noise ratio [21]: The classical two-state Gilbert-Elliott model [4, 6] for burst noise channels, which characterizes error sequences, has been widely used and analyzed. In [17], a multiple-state quasi-stationary Markov channel model is used to characterize the wireless nonstationary channel. In [18, 22], a finite-state Markov channel was described that has multiple states representing the reception at different signal-to-noise levels. A fluid version of the Gilbert-Elliott model was used in [11] to perform analysis of delay and packet-discard performance as well as the effective capacity for QoS support over a wireless link with ARQ and FEC. In the following, we will describe a very general framework to analyze delay peformance on wireless links. We will illustrate it with the example of a Rayleigh fading channel model with hybrid ARQ/FEC error control. The wireless link will be modeled as a fluid version of Finite-State Markov model. It is important to note that the framework presented here is by far not limited to this particular channel (Rayleigh with ARQ/FEC), but can be applied to many other models as well.

In addition to a model for underlying wireless links, in order to provide real-time guarantees, one needs an appropriate description of the workload carried on links: the traffic model. This model, in turn, depends upon the desired service requirements. Real-Time communication service requirements can be guaranteed in two forms: deterministic services and statistical services. *Deterministic services* require that the delay and delivery guarantees are satisfied for all packets, and tend to heavily overcommit resources. *Statistical services* allow packets to be occasionally dropped or excessively delayed. Statistical services thus significantly increase the efficiency of network usage by allowing increased statistical multiplexing of the underlying network resources. A number of approaches have been presented in the literature to provide statistical guarantees for determin-

Shengquan Wang, Ripal Nathuji, Riccardo Bettati and Wei Zhao are with the department of Computer Science, Texas A&M University, College Station, TX 77843. Email: {swang,rnathuji,bettati,zhao}@cs.tamu.edu.

istically constrained traffic streams. We will make use of *rate-variance envelopes* [10].

In the following, we focus on providing end-to-end delay guarantees via connection admission control mechanisms that make sure that end-to-end delay requirements are not violated both for new and existing connections after a new connection has been admitted. We adopt two different admission control mechanisms that differ from each other by the point in time during which explicit delay computations are performed. We first introduce Delay-Based Admission Control (DBAC), an approach that performs delay tests at connection establishment time. The second scheme is a Utilization-Based Admission Control (UBAC) mechanism. In UBAC, offline computations are used to perform a simple utilization-based test along the path of an incoming flow. We will show in our experimental evaluation that UBAC's performance in terms of admission probability is comparable to that of the much more expensive DBAC scheme.

2. Models of Wireless Networks and Links

2.1. Overview

We consider a wireless network that consists of a number of wireless links, each of which connects two wireless nodes. Fig. 1 shows an example of a wireless system that falls into our network model.

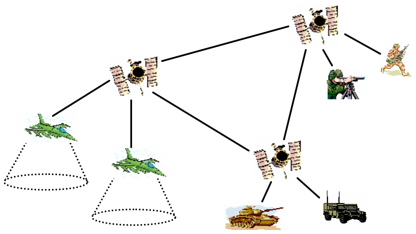


Figure 1. An Ground-space-ground Wireless Communication System

To guarantee an end-to-end delay, delay characteristics on each wireless link need to be analyzed. Underlying wireless links are physical *wireless channels*. For the purpose of delay guarantees, a wireless channel model describes the channel error statistics and the effect the latter has on channel capacity. A large number of such models have been described and evaluated in the literature, based on the Rayleigh Fading Channel, or (by adding a line-of-sight component) the Rician Fading Channel [15]. Typical channel error statistics models, such as the binary symmetric channel, are modeled as finite-state Markov models, and can be used to represent time-varying Rician (and others) channels in a variety of settings [1, 7, 8].

The formulation of a *link model* has to account for error control schemes used at the link layer. In this section, we will largely follow the approach presented by Krunz and Kim in [11] to map channel and error-correction schemes into a Markov link model. We will extend their two-state Markov model to a more general finite-state Markov model that we will use to derive the stochastic service curve to perform delay analysis later. In the following, we first consider the framework of the wireless link, and then lay out a more detailed description of our Markov link model.

2.2. Framework of a Wireless Link

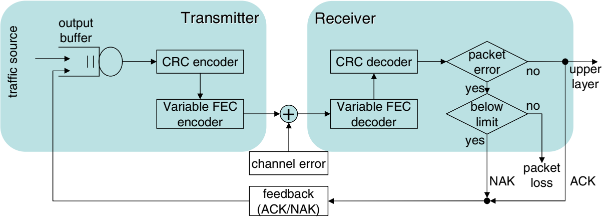


Figure 2. Wireless Link Framework [11]

We consider a hybrid ARQ/FEC error control scheme (Fig. 2) and assume a *stop-and-wait* (SW) scheme for ARQ. The FEC capability in the hybrid ARQ/FEC mechanism is characterized by three parameters: the number of bits in a code block (n), the number of payload bits (k), and the maximum number of correctable bits in a code block (r).The FEC code rate e is defined as $e = \frac{k}{n}$. Assuming that a FEC code can correct up to r bits and that bit errors in a given channel state are independent, the probability P_{nc} that a packet contains a non-correctable error, given a bit error rate p, is given by [11]

$$P_{nc}(p) = \sum_{j=r+1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) p^j (1-p)^{n-j}. \qquad (1)$$

To account for the FEC overhead, the actual average service capacity C_e observed at the output of the buffer is $C_e = C \cdot e$, where C is the maximum capacity for the wireless channel.

2.3. Markov Link Model

Though statistical characteristics of a wireless channel can significantly vary with time, the basic system parameters remain constant over short time intervals. Thus it can be modeled as a quasi-stationary channel. We use a fluid version of a finite-state Markov-Modulated model with L

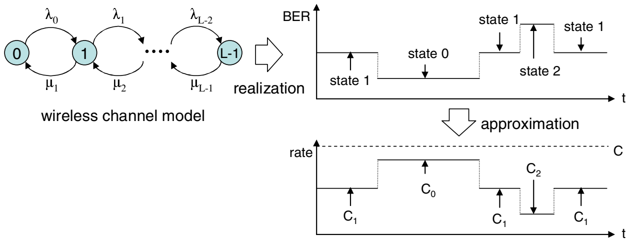


Figure 3. Fluid Version of Finite-State Markov Model of a Wireless Channel

states $(0, 1, \ldots, L-1)$ as shown in Fig. 3 [11]. The bit error rates (BER) during State i are given by p_i, where we assume $0 \leq p_0 < p_1 < \cdots < p_{L-1} \leq 1$. The durations in State i before being transitioned to State $i+1$ and $i-1$ are exponentially distributed with mean $\frac{1}{\lambda_i}$ and $\frac{1}{\mu_i}$, respectively. It is typically assumed that the transitions only happen between adjacent states.

It is generally difficult to get analytically tractable results that accurately represent the behavior of ARQ and FEC and so accurately map the channel model into the respective link model. To solve this, the authors in [11] assume that the packet departure process follows a fluid process with an average constant service capacity that is modulated by the channel state (Fig. 3). Each state i then gives raise to a stationary link-layer service capacity C_i, which can be determined as follows: A packet is repeatedly retransmitted until it is correctly received at the destination. Let N_{tr} denote the number of retransmissions (including the first transmission) until a packet is successfully received. For a given packet error probability p_i of the channel in State i, the expected value of N_{tr} is $E[N_{tr}] = \frac{1}{1-p_i}$. Thus, C_i can be written as [11]

$$C_i = C \cdot e \cdot (1 - P_{nc}(p_i)). \tag{2}$$

As the state transition rates of the wireless-channel model are not affected by ARQ or FEC, the result is a Markov-modulated model with L State $(0, 1, \ldots, L-1)$ each associated with capacity C_i.

2.4. Stochastic Service Curve of a Wireless Link

In order to determine the performance guarantees that can be given by a wireless link, we must describe the amount of service that the link can provide. For this we make use of so-called *service curves*. The *stochastic service curve* $S(t) = \int_0^t C(\tau)d\tau$ is defined as the traffic amount that can be served during time interval $[0, t]$ by the wireless channel, where $C(\tau)$ is the capacity at time τ. Correspondingly, we define $S_i(t)$ as the traffic amount that can be served during time interval $[0, t]$ with the system in State i at time t, $F_i(t, x)$ and $F_S(t, x)$ as the cumulative probability distribution of $S_i(t)$ and $S(t)$, respectively. We denote π_i as the probability that the link is in State i at any time when the system is steady, and we then have

$$F_S(t, x) = \sum\nolimits_{l=0}^{L-1} \pi_i F_i(t, x). \tag{3}$$

We need to compute $F_i(t, x)$: Following a standard fluid approach [16], we proceed by setting up a generating equation for $F_i(t, x)$ at an incremental time Δt later in terms of the probabilities at time t.

$$\begin{aligned} F_i(t + \Delta t, x) \quad = \quad & (\lambda_{i-1}\Delta t)F_{i-1}(t, x - C_{i-1}\Delta t) \\ & +(1 - (\mu_i + \lambda_i)\Delta t)F_i(t, x - C_i\Delta t) \\ & +(\mu_{i+1}\Delta t)F_{i+1}(t, x - C_{i+1}\Delta t), \end{aligned} \tag{4}$$

as $i = 1, \ldots, L-2$, and

$$\begin{aligned} F_0(t + \Delta t, x) \quad &= \quad ((1 - \lambda_1)\Delta t)F_0(t, x - C_0\Delta t) \\ & \quad\quad +(\mu_1\Delta t)F_1(t, x - C_1\Delta t), \end{aligned} \tag{5}$$

$$\begin{aligned} F_{L-1}(t + \Delta t, x) \quad &= \quad (\lambda_{L-2}\Delta t)F_{L-2}(t, x - C_{L-2}\Delta t) \\ & \quad\quad +((1 - \mu_{L-1})\Delta t) \\ & \quad\quad F_{L-1}(t, x - C_{L-1}\Delta t). \end{aligned} \tag{6}$$

Both sides are divided by Δt in the above equations, and as $\Delta t \to 0$, with some algebraic manupulation, the above equations become partial differential equations, which can be rewritten in matrix form as follows:

$$\frac{\partial \mathbf{F}}{\partial t} + \mathbf{C}\frac{\partial \mathbf{F}}{\partial x} \quad = \quad \mathbf{QF}, \tag{7}$$

where $\mathbf{F} = (F_0(t, x), F_1(t, x), \ldots, F_{L-1}(t, x))^{\perp}$, $\mathbf{C} = \mathrm{diag}(C_0, C_1, \ldots, C_{L-1})$ and

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \mu_1 & \cdots & 0 \\ \lambda_0 & -(\lambda_0 + \mu_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mu_{L-1} \\ 0 & 0 & \cdots & -\mu_{L-1} \end{pmatrix}. \tag{8}$$

The initial conditions are $F_i(0, x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$ for $i = 0, 1, \ldots, L-1$.

The above linear first-order hyperbolic PDEs can be solved numerically, and the $F_i(t, x)$'s can be computed. Furthermore, if we define $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_{L-1})^{\perp}$, the π_i's in (3) are given by

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{Q}, \text{ and } |\boldsymbol{\pi}| = 1. \tag{9}$$

Fig. 4 shows simulated data for the distribution of $S(t)$ for a two-state Markov model, where we specify $C = 2$ Mbps, $t = 0.5$ s, $\lambda_0 = 10, \lambda_1 = 30, p_0 = 10^{-6}$, and we vary the BER p_1 and the code parameters (n, k, r). The data illustrates that BER and coding substantially affect the service distribution.

In Section 4, we will illustrate how the service distribution $F_S(t, x)$ is used to perform statistical delay analysis.

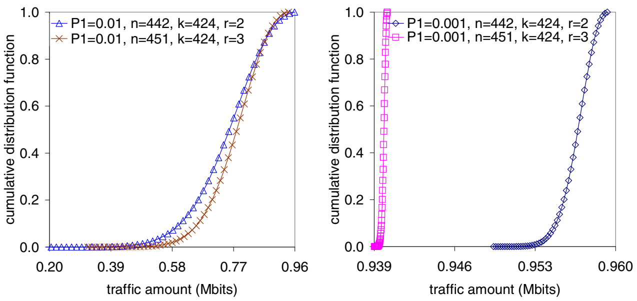


Figure 4. The Stochastic Service Curve for a Wireless Link

3. Traffic Model

We model the *traffic arrival* for a flow as a stochastic arrival process $\mathcal{A} = \{\mathcal{A}(\tau), \tau \geq 0\}$, where random variable $\mathcal{A}(\tau)$ denotes the incoming traffic amount of the flow for a link server during time interval $[0, \tau]$. The arrival process $\mathcal{A}$ is stationary and ergodic. Since $\mathcal{A}(\tau)$ is stationary, $\mathcal{A}(\tau + t) - \mathcal{A}(\tau)$ possesses the same probability distributions for all τ. Therefore, we can define a random variable $R(t) = \frac{\mathcal{A}(t_0+t)-\mathcal{A}(t_0)}{t}$ as the *stochastic traffic arrival rate*. The traffic arrival can be bounded either deterministically or stochastically by the traffic arrival envelope as follows:

Definition 3.1 (Deterministic Traffic Arrival Envelope) *The function $b(t)$ is called the deterministic traffic arrival envelope of the traffic arrival with rate $R(t)$ if*

$$\int_{t_0}^{t_0+t} R(\tau)d\tau \leq b(t), \tag{10}$$

for any $t_0, t \geq 0$.

For example, the traffic arrival can be constrained by a leaky bucket with parameters (σ, ρ) as $\int_{t_0}^{t_0+t} R(\tau)d\tau \leq \sigma + \rho \cdot t$, for any $t_0, t \geq 0$, where σ is the burst size and ρ is the average rate.

Definition 3.2 (Statistical Traffic Arrival Envelope) *The distribution $B(t)$ forms the statistical traffic arrival envelope of the traffic arrival $\mathcal{A}$ if*

$$\int_{t_0}^{t_0+t} R(\tau)d\tau \preceq_{st} B(t), \tag{11}$$

for any $t_0, t \geq 0$, where $X \preceq_{st} Y$ means $P\{X < Z\} \leq P\{Y < Z\}$.

We will be describing traffic arrival using the rate-variance envelope [10]. It is a key factor for computing delay violation probabilities.

4. Statistical Delay Analysis in a Wireless Network

A probabilistic real-time guarantee can be defined as a bound on the probability of exceeding a deadline, i.e., $P\{D > d\} \leq \epsilon$, where the delay D suffered by a packet is a random variable, d is the given deadline, and ϵ is the given violation probability (generally small).

We consider networks that use static-priority schedulers at the network nodes, as opposed to previous work considering FIFO buffers [11]. For *wired* networks with static priority scheduling, we addressed the issue of how to provide statistical real-time guarantees in [19], based on Knightly's earlier work in [10]. Define C as the capacity of a link and G_i as a group of flows that are served by the link at priority i. Assume $b_{i,j}(t)$ and $B_{i,j}(t)$ to be the deterministic and statistical bound, respectively, for the traffic arrival for the individual flow $j \in G_i$. Then the *delay violation probability* $\Pr\{D_i \geq d_i\}$ for a random packet with priority i at the output link can be bounded by

$$\Pr\{D_i \geq d_i\} \leq \max_{t<\beta_i} \Pr\{B^*(t+d_i) \geq C \cdot (t+d_i)\}, \tag{12}$$

where $B^*(\cdot)$ is the amount of aggregated traffic of same and higher priorities:

$$B^*(t+d_i) = \sum_{q=1}^{i-1} \sum_{j \in G_q} B_{q,j}(t+d_i) + \sum_{j \in G_i} B_{i,j}(t), \tag{13}$$

and β_i is a bound on the busy interval defined as follows:

$$\beta_i = \min\{t > 0 : \sum_{q=1}^{i} \sum_{j \in G_q} b_{q,j}(t) \geq C \cdot t\}. \tag{14}$$

The above formula cannot be applied directly for wireless links however, as their capacity varies over time. Fortunately, as the following observation shows, it is not difficult to integrate stochastic arrivals and a stochastic service curve to compute delay violation probabilities: Consider a wireless link with a static-priority scheduler and maximum capacity C. Let $C(t)$ be the available capacity for traffic as a function of time. Thus $C - C(t)$ is the *unavailable* capacity of link at time t. We can equivalently model this system if we define a *virtual traffic arrival* with instantaneous capacity $C - C(t)$ to a link with constant capacity C, by requiring that this virtual traffic is given strictly highest priority during scheduling. Packet delays for real traffic in the original system are identical to delays in this virtual-traffic model. In particular, if the wireless link has a stochastic service curve $S(t)$, then the equivalent virtual traffic on the wireless link has the stochastic envelope $B'(t) = C \cdot t - S(t)$. This gives raise to the following theorem:

Theorem 4.1 *Consider a wireless link with a static-priority scheduler and stochastic service curve $S(t)$. Assume $B_{i,j}(t)$*

is the statistical bound for the traffic arrival of the individual flow $j \in G_i$. Then, the delay violation probability for a random packet with priority i can be bounded by

$$\begin{aligned}\Pr\{D_i \geq d_i\} \leq \max_{t>0} \Pr\{B'(t+d_i) \\ +B^*(t+d_i) \geq C \cdot (t+d_i)\},\end{aligned} \qquad (15)$$

where $B'(t) = C \cdot t - S(t)$, and $B^*(t+d_i)$ is defined in (13). [1]

We make the following observations about Theorem 4.1: First, $B'(t+d_i)$ and $B^*(t+d_i)$ are independent. Given their distribution functions, the distribution function of the summation $B^*(t+d_i)+B'(t+d_i)$ can be obtained by their direct convolution. Second, the distribution of $B'(t+d_i)$ can be directly obtained from $S(t)$, which we in turn derived in Section 2. Note that (15) holds for any $S(t)$, no matter what specific wireless link model is chosen. The main challenge for statistical delay analysis is how to obtain the distribution function of $B'(t+d_i)$, i.e., how to clearly describe the traffic arrival envelope. It is inherently very difficult for the network to enforce or police the stochastic properties of traffic streams. Consequently, if a particular application does not conform to the chosen stochastic model, no guarantees can be made. Moreover, if admitted to the network, such a non-conforming stream could adversely affect the performance of other applications if it is statistically multiplexed with them. Therfore, we must find a means to describe the non-conforming traffic so that we can perform delay analysis.

We will use the approach previously developed in [19] for the statistical delay analysis. We represent the input traffic flows as a set of random processes. Traffic policing ensures that these processes are independent. If we know the mean value and the variance of each individual traffic random variable, and the number of flows is large enough, then by the *Central Limit Theorem*, we can approximate the random process of the set of all flows combined. The *Central Limit Theorem* states that the summation of a set of independent random variables converges in distribution to a random variable that has a *Normal Distribution*. In the following, we illustrate how using *rate-variance envelopes*, the mean rate and the rate-variance of each individual flow can be determined by deterministic traffic models.

The rate-variance envelope $RV(t) = var(R(t))$ describes the variance of the arrival rate for the incoming flow over an interval of length t [10]. We assume that a flow of priority i is controlled by a leaky bucket with burst size σ_i and average rate ρ_i at each router. Assume that flow j in the group of flows G_i has mean rate $\phi_{i,j}$ and rate-variance envelope $RV_{i,j}(t)$. With application of a Gaussian approximation over intervals, $B^*(t+d_i)$ in (15) can be approximated by a normal distribution $N(\phi_i(t), RV_i(t))$ [10], where

$$\phi_i(t) = (t+d_i)\sum_{q=1}^{i-1}\sum_{j\in G_q}\phi_{q,j} + t\sum_{j\in G_i}\phi_{i,j}, \qquad (16)$$

$$\begin{aligned} RV_i(t) &= (t+d_i)^2 \sum_{q=1}^{i-1}\sum_{j\in G_q} RV_{q,j}(t+d_i) \\ &\quad + t^2 \sum_{j\in G_i} RV_{i,j}(t). \end{aligned} \qquad (17)$$

Given the deterministic traffic arrival envelope $b_{i,j}(t) = \sigma_i + \rho_i t$, for any flow j in G_i, we can easily obtain mean rate $\phi_{i,j}$ for each individual flow, and an adversarial mode is chosen for obtaining the rate-variance envelope $RV_{i,j}(t)$ [10]. [2] We obtain the *mean rate* and the *rate-variance envelope* as follows:

$$\phi_{i,j} = \rho_i, \qquad (18)$$

$$RV_{i,j}(t) \leq \frac{\rho_i \sigma_i}{t}. \qquad (19)$$

In summary, this leads to the following lemma:

Lemma 4.2 *Define $n_q = |G_q|$, $q = 1, 2, \ldots, i$. With application of a Gaussian approximation over intervals, $B^*(t+d_i)$ can be bounded by a normal distribution $N(\phi_i(t), RV_i(t))$, i.e.,*

$$\Pr\{B^*(t+d_i) < x\} \leq \mathbf{\Phi}(\frac{x-\phi_i(t)}{\sqrt{RV_i(t)}}), \qquad (20)$$

where

$$\phi_i(t) = (t+d_i)\sum_{q=1}^{i-1} n_q\rho_q + tn_i\rho_i, \qquad (21)$$

$$RV_i(t) \leq (t+d_i)\sum_{q=1}^{i-1} n_q\rho_q\sigma_q + tn_i\rho_i\sigma_i, \qquad (22)$$

$$\mathbf{\Phi}(a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}\exp(-\frac{x^2}{2})dx. \qquad (23)$$

The distribution function of the summation $B^*(t+d_i)+B'(t+d_i)$ can be obtained by convolution. Define this distribution function as $F_B(t+d_i, x)$. Then, the delay violation probability can be upper-bounded with utilization as shown in the following theorem:

Theorem 4.3 *Consider a wireless link with a static-priority scheduler and the stochastic service curve $S(t)$. Assume the same traffic envelope in Theorem 4.1. The delay violation probability for any packet with priority i is bounded by*

$$\Pr\{D_i \geq d_i\} \leq 1 - \min_{t>0} F_B(t+d_i, C\cdot(t+d_i)). \qquad (24)$$

[1] Here the maximum busy interval is canceled out due to the possibly unconstrained stochastic service curve. The virtual traffic may produce infinite-length maximum busy interval. So the delay violation probability may appear to be loose. In our simulation data, we find that the maximum value will be achieved for relatively small values of t, therefore, the bound is tight.

[2] In adversarial mode, the traffic arrival process conforms to a binomial distribution, where the rate-variance envelope is upper bounded.

Having derived the statistical delay formula for a single wireless link, we obtain the end-to-end delay violation probability along each path. Given the delay violation probability ϵ_i and the end-to-end deadline d_i along route $\mathcal{R}$, we can partition d_i into $\{d_i^k : k \in \mathcal{R}\}$, and the delay guarantee is met when [9]

$$\begin{aligned} & P\{D_i^{e2e} > \sum\nolimits_{k\in\mathcal{R}} d_i^k\} \\ \leq & \; 1 - \prod\nolimits_{k\in\mathcal{R}}(1 - P\{D_i^k > d_i^k\}) \qquad (25) \\ \leq & \; \epsilon_i. \qquad (26) \end{aligned}$$

This bound on the end-to-end real-time guarantee gives raise to several possible approaches to admission control and connection establishment.

5. Admission Control Mechanisms

Recall that it is the admission control mechanism that decides if a new connection can be admitted. The decision is based upon whether the end-to-end delay requirements can be met for both newly arriving and existing connections. In this section, we describe two approaches to admission control that perform the delay computation previously described.

5.1. Delay-Based Admission Control (DBAC)

Delay-based admission control (DBAC) is a mechanism that makes admission decisions by analyzing the system state at run-time. In particular, DBAC performs delay calculations at flow establishment time using run-time flow information to determine both whether the introduction of the new flow allows existing flows to maintain their real-time requirements and whether guarantees can be provided for the new flow. Note that in systems with multiple priorities, only flows of equal and lower priorities need to be checked.

We employ the equations from previous sections to calculate the delay violation probabilities required in the DBAC algorithm. Though the DBAC algorithm is conceptually simple, it is computationally complex since it requires solving systems of partial differential equations (7) and performing convolution operations in (24). It is also necessary for the admission controller to be aware of the network topology and routing. Thus DBAC becomes extremely costly as the system scales in size. [3]

5.2. Utilization-Based Admission Control (UBAC)

With UBAC, we decrease admission control overhead by reducing the amount of computation that needs to be performed at run-time. For this to be possible, we assume that each link server reserves a certain percentage of capacity for every particular traffic priority class. It is the responsibility of the admission control module to ensure that the capacity usage of individual priority classes do not exceed the reserved portion. This is necessary to provide isolation among different priority traffic and hence to guarantee end-to-end delays to the flows.

As opposed to run-time calculations per flow, UBAC requires off-line delay computations per priority traffic to obtain what we call a safe utilization bound. Since flow population information is unavailable for off-line calculations, we must obtain a *flow-population-insensitive* statistical delay formula. During run-time, UBAC checks whether the link utilization allocated to each priority traffic is not exceeded. The total number n_i of flows of priority i on a link is therefore subject to the following constraint:

$$n_i \leq \frac{\alpha_i}{\rho_i} C, \qquad (27)$$

where α_i is the ratio of the link capacity allocated to traffic of priority i and ρ_i is the average rate of priority i traffic. With this constraint, the mean rate and the rate-variance can be upper-bounded as follows:

$$\begin{aligned} \phi_i(t) &= (t + d_i)\sum\nolimits_{q=1}^{i-1} \alpha_q C + t\alpha_i C, \qquad (28) \\ RV_i(t) &= (t + d_i)\sum\nolimits_{q=1}^{i-1} \alpha_q \sigma_q C + t\alpha_i \sigma_i C. \qquad (29) \end{aligned}$$

Correspondingly, Lemma 4.2, Theorem 4.3 and Equation (26) can be re-formulated using the flow-population insensitive definition for the new $\phi_i(t)$ and $RV_i(t)$ given above. Thus we observe that the benefit of UBAC over DBAC is its ability to perform admission control without heavy run-time computations. As our performance evaluation will illustrate, UBAC is still able to provide comparable system efficiency.

6. Performance Evaluation

In this section, we evaluate the performance of the two approaches discussed in the previous sections. The simulated wireless network could be representative of a ground-space-ground wireless communication system (Fig. 1). We allow any pair of nodes in the network to establish a real-time priority connection.In our wireless link model, we assume that all links in the network have a maximum capacity of 2 Mbps. Links follow a two-state Markov model as previously defined. In the simulation, we specify the link parameters as follows: $\lambda_0 = 10, \lambda_1 = 30, p_0 = 10^{-6}$, and we vary the bit error rate (BER) p_1 for State 1 (BAD state). We also adopt five different Bose-Chaudhuri-Hocquenghem (BCH) [14] coding schemes for FEC. We assume that requests for real-time flow establishment form a Poisson process, and

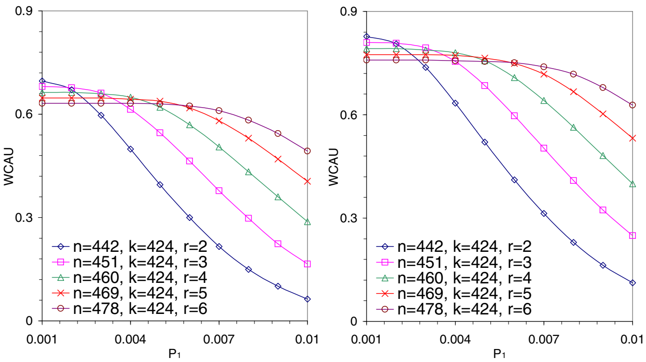
[3] Note that the overhead only occurs at flow establishment time, not during packet forwarding time.

that flow lifetimes are exponentially distributed with an average of 180 seconds. [4]

In obtaining our results, we are interested in two metrics: i) *WCAU* – The *worst-case achievable utilization* is the maximum link utilization that can be safely allocated to real-time traffic in UBAC; ii) *Admission Probability* – This is the probability that a flow can be admitted without violating delay guarantees. Both metrics reflect on the efficient use of network resources.

6.1. WCAU Comparison

The underlying network topology in the WCAU experiment is the network shown in Fig. 1, where nodes communicate through a space-based reach-back network. We vary the link characteristics by varying the bit error rate (BER) p_1 for State 1 (BAD state). We also consider five different BCH coding schemes with increasing level of correctability (i.e., different (n, k, r) [11]. In our traffic model, we assume that all traffic belongs to a single real-time priority. We simulate voice traffic, with bursts $\sigma = 640$ bits, and average rate $\rho = 32000$ bps. We assume that the end-to-end deadline is 15 ms. The end-to-end deadline violation probability is either 10^{-6} or 10^{-3}.



(a) $\epsilon = 10^{-6}$ (b) $\epsilon = 10^{-3}$

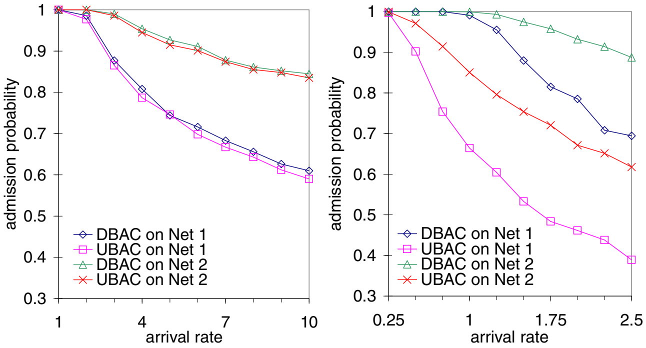
Figure 5. WCAU Comparison

The WCAU can be computed by (26) we obtained in the previous section using simple binary search. The results of our WCAU experiments are shown in Fig. 5. The following observations can be made from these results: 1). *Sensitivity of WCAU to channel coding:* Our results show the performance tradeoff of using various channel codes. Codes which provide greater error correction decrease the amount of actual traffic included in packets. For low error rates, this capability is not worthwhile, as shown in Fig. 5, since error correction is rarely useful, and thus decreases the overall achievable utilization. 2). *Sensitivity of WCAU to BER:* As the BAD-state BER p_1 is increased from 0.001 to 0.01, the WCAU decreases for all cases. These results support the intuition that, as the error probability of the network increases, the amount of capacity that can be supported for real-time traffic should decrease. 3). *Sensitivity of WCAU to deadline violation probability:* As expected, the WCAU increases when the deadline violation probability is decreased. In other words, allowing higher loss probabilities creates additional available utilization for real-time traffic.

6.2. Admission Probability Comparison

In addition to the topology (Fig. 6) used in the previous section (called Net 1 in this content), we utilize a randomly generated topology with the same number of total nodes, referred to as Net 2. We use this randomly generated topology in order to support the fact that our results are not dependent upon a particular topology. We fix the bit error rate (BER) p_1 for State 1 (BAD state) $p_1 = 0.001$ and choose BCH coding scheme with parameters $(n = 442, k = 424, r = 2)$. The end-to-end deadline violation probability is 10^{-6}.

We simulate the case that there is only a single real-time priority in the network with the parameters σ, ρ, d equal to the values used in the first simulation. We also simulate the case that there are two real-time priorities in the network to see how multiple priorities affect the admission probability. In this case, we choose additional higher-priority traffic as follows: $\sigma = 1280$ bits, $\rho = 64000$ bps, $d = 0.005$ s. The capacity is allocation with ratio $\alpha_{high} : \alpha_{low} = 1 : 3$.



(a) single priority (b) two priorities

Figure 6. Admission Probability Comparison

As expected, in both cases, as the flow arrival rate increases, admission probability decreases. The substantial conclusion we draw from these results is with regard to the relationship between UBAC and DBAC: It is clear from

[4] A real system would support best-effort traffic as well. Since this traffic would not affect the results of this evaluation, we omit it from our experiments.

Fig. 6(a) that with the single priority case, UBAC is in fact able to provide the same efficiency with regard to network resource allocation as DBAC. This result is significant because it means that the efficiency of DBAC can be provided with low run-time overhead by using UBAC. Thus costly run-time delay computations can be removed without sacrificing performance. From Fig. 6(b), we find that DBAC obtains more gains in terms of admission probability than UBAC when multiple priority classes are used. This can be attributed to the fact that the pre-allocation of capacity in UBAC disable the capacity sharing between the different priorities traffic so that the overall achievable utilization is decreased. Therefore, DBAC achieves much higher admission probability than UBAC in the multiple-priority case.

7. Conclusions

The statistical nature of service provided by wireless links inherently precludes deterministic delay guarantees. Means must therefore be provided that allow definition and enforcement of *statistical* guarantees. We present *statistical service curves* and show how they accurately represent the service provided by wireless links in a tractable manner. Statistical traffic arrival descriptions, which should be used due to the stochastic nature of service, are very impractical to police. So we are studying to use methods that accurately capture the statistical behavior of deterministically bounded traffic.

In this paper, we described how statistical service curves could be applied for static-priority schedulers using what we call "virtual traffic" to compute the available service to real-time traffic. We did not assume a particular traffic pattern; instead we used rate-variance envelopes, a simple and general traffic characterization. This methodology made our approach applicable to any particular situation. We also evaluated two different admission control mechanisms (delay-based and utilization-based) and illustrate performance trade-offs between resource utilization and overhead.

References

[1] P. Bello. Aeronautical channel characterization. *IEEE Transactions on Communications*, 21(5):548–563, May 1973.

[2] J. B. Cain and D. N. McGregor. A recommended error control architecture for ATM networks with wireless links. *IEEE Journal on Selected Areas in Communications*, 15(1):16–28, Jan 1997.

[3] A. Dailianas and A. Bovopoulis. Real-time admission control algorithms with delay and loss guarantees in ATM networks. In *Proceedings of IEEE Infocom*, Toronto, Canada, June 1994.

[4] E. O. Elliott. Estimates of error rates for codes on burst-noise channels. *Bell Syst. Tech. J.*, 42(9):1977–1997, Sept 1963.

[5] R. Fantacci. Queuing analysis of the selective repeat automatic repeat request protocol wireless packet networks. *IEEE Transactions on Vehicular Technology*, 45(2):258–264, May 1996.

[6] E. N. Gilbert. Capacity of a burst-noise channel. *Bell Syst. Tech. J.*, 39(8):1253–1265, Sept 1960.

[7] J. Hagenauer and W. Papke. Data transmission for maritime and land mobile using stored channel simulation. In *Proceeding of IEEE Veh. Technol. Conference*, San Diego, CA, USA, 1982.

[8] R. W. Huck, J. S. Butterworth, and E. E. Matt. Propagation measurements for land mobile satellite services. In *Proceeding of IEEE Veh. Technol. Conference*, Toronto, Canada, 1983.

[9] E. Knightly. H-bind: A new approach to providing statistical performance guarantees to VBR traffic. In *Proceedings of IEEE INFOCOM*, San Francisco, CA, USA, Mar 1996.

[10] E. Knightly. Enforceable quality of service guarantees for bursty traffic streams. In *Proceedings of IEEE Infocom*, San Francisco, CA, USA, March 1998.

[11] M. Krunz and J. G. Kim. Fluid analysis of delay and packet discard performance for QoS support in wireless networks. *IEEE Journal on Selected Areas in Communications*, 19(2):384–395, Feb 2001.

[12] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proceedings of ACM Sigmetrics*, Newport, RI, June 1992.

[13] J. Liebeherr, D. Wrege, and D. Ferrari. Exact admission control in networks with bounded delay services. *IEEE/ACM Transactions on Networking*, 4(6):885–901, December 1996.

[14] S. Lin and J. D. J. Costello. *Error Control Coding: Fundamentals and Applications*. Prentice Hall, New Jersey, 1983.

[15] J. G. Proakis. *Digital Communications*. McGraw-Hill, New York, 1995.

[16] M. Schwartz. *Broadband Integrated Networks*. Prentice Hall, New Jersey, 1996.

[17] B. Vucetic. An adaptive coding scheme for time-varying channels. *IEEE Trans. Commun.*, 39:653–663, May 1991.

[18] H. S. Wang and N. Moayeri. Finite-state Markov channel - a useful model for radio communication channels. *IEEE Transactions on Vehicular Technology*, 45(2):258–264, May 1996.

[19] S. Wang, D. Xuan, R. Bettati, and W. Zhao. Differentiated services with statistical real-time guarantees in static-priority scheduling networks. In *Proceedings of IEEE RTSS*, London, UK, December 2001.

[20] S. Wang, D. Xuan, R. Bettati, and W. Zhao. Providing absolute differentiated services for real-time applications in static-priority scheduling networks. In *Proceedings of IEEE Infocom*, Anchorage, Alaska, USA, April 2001.

[21] D. Wu and R. Negi. A wireless channel model for support of quality of service. In *Proceedings of IEEE GLOBECOM*, San Antonio, TX, USA, Nov 2001.

[22] Q. Zhang and S. A. Kassam. Finite-state Markov model for Rayleigh fading channels. *IEEE Transactions on Communications*, 47(11):1688–1692, Nov 1999.