RED Theory for QoS Provisioning in Wireless Communications

Xihua Dong, Xiaochen Li and Dapeng Wu

Abstract

In this paper, we study the performance limit of a wireless communication system over a fading channel. The system under study consists of 1) a finite-buffer discrete-time queueing system on the link layer, and 2) a rate-adaptive channel coding system on the physical layer. The objective of this paper is to analyze the relationship among data rate (R), packet error probability (E), and delay bound (D), under the interaction between the link layer and the physical layer. In our analysis, we consider three types of packet errors, i.e., 1) packet drop due to full buffer, 2) packet drop due to delay bound violation, and 3) packet decoding error due to channel noise. We obtain an upper bound on the packet error probability. Furthermore, by minimizing the packet error probability over the transmission rate, we obtain an optimal rate control policy that guarantees the user-specified data rate and delay bound. In the case of constant arrival, the optimal rate control policy results in an RED triplet; then by varying data rate and delay bound, we obtain RED Pareto-optimal surface, which serves as the performance limit of the system under study.

Index Terms

Channel capacity, fading channel, optimal rate control policy, RED, QoS.

I. Introduction

Future wireless networks are targeted at supporting various applications such as voice, data, and multimedia over packet-switched networks. Many of these applications require quality of service (QoS) guarantees, e.g., data rate, packet error probability, and delay bound. However, fading in wireless channels may cause severe QoS violations. Hence, providing QoS guarantees poses a great challenge for the design of next-generation wireless networks.

Data communication over fading channels without delay constraint has been extensively studied in the literature. When the delay constraint is absent, the maximum expected throughput is Shannon's ergodic capacity. Depending on the assumptions of channel state information (CSI) availability at the transmitter (CSIT) and at the receiver (CSIR), existing works can be categorized by the following categories: 1) channels with perfect CSIT and CSIR [1], 2) the finite-state Markov channels (FSMC) without CSI [2], 3) channels where CSIT is a deterministic function of CSIR [3], and 4) channels with causal CSI [4], among other variants. Some works have also addressed a more realistic case of non-perfect CSI [5],

The authors are with Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611. Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu, http://www.wu.ece.ufl.edu. This work was supported in part by an Intel gift, the US National Science Foundation under grant DBI-0529012 and CNS-0643731, and the US Office of Naval Research under grant N000140810873.

[6]. Practical adaptive modulation and coding schemes for data communication over fading channels are studied in [7]–[9]. Outage capacity is another widely-studied capacity notion for a fading channel.

The aforementioned works do not consider queueing delay, and thus are not applicable to QoS provisioning for delay sensitive applications. Recently, delay-constrained communication has received a lot of attention. In [10], Hanly and Tse proposed the concept of delay-limited capacity, which is defined as the maximum achievable rate achievable under the constraint of a delay bound and zero delay-bound-violation probability. Throughput maximization under average delay constraint is studied in [11]–[15]. However, these works have two limitations in QoS provisioning: 1) the average delay constraint cannot specify delay bound violation probability, needed by many delay-sensitive applications, e.g., interactive games and real-time video; 2) these works assume that the buffer size is infinite while practical systems have limited buffer space. This motivates us to investigate statistical delay guarantee (i.e., guarantee on delay bound violation probability) for wireless communication systems with finite buffer.

In this work, we consider a single-user communication system with finite buffer. Packet transmission is subject to a constraint on delay bound violation probability. In addition, different from most existing works that address transmission delay, e.g., [11], [12], we consider non-perfect channel coding, which incurs non-zero decoding error probability. Thus a packet under our study may experience three types of errors: 1) delay bound violation, 2) packet drop due to full buffer, and 3) decoding error due to channel noise. In this work, we develop an upper bound on the packet error probability that characterizes the union of the events of packet drop, delay bound violation, and incorrect decoding. Since we consider a finite buffer, our result holds for arbitrary delay bound, i.e., our analysis holds for both the small delay regime and the large delay regime. Based on our analytical result for packet error probability, we obtain an optimal rate control policy that guarantees the user-specified QoS, by minimizing the packet error probability over the transmission rate. The optimal rate control policy results in an RED (rate-error-delay) triplet; then in the case of constant arrival, by varying data rate and delay bound, we obtain RED Pareto-optimal surface, which serves as the performance limit of the system under study. The RED Pareto-optimal surface serves the same role as Shannon's channel capacity, i.e., it tells how far away a practical system (consisting of link/physical layers) is from the optimal performance. In addition, our results provide important insights about optimal rate control policy for joint link layer and physical layer design.

This paper is closely related to Ref. [16], which studies the optimal rate and power control policies for maximizing error-free throughput under the condition of buffer overflow and bit errors in the physical layer. Different from their work, this paper studies the fundamental tradeoff among data rate, packet error probability, and delay bound, and provides explicit relationship among data rate, packet error probability, and delay bound, which is characterized by RED Pareto-optimal surface; in other words, Ref. [16] studies a single point on our RED Pareto-optimal surface, and Ref. [16] does not study how the maximum rate changes with packet error probability and delay bound. Packet drop and decoding error are also addressed in Refs. [17], [18]. One of the major differences between our work and Refs. [17], [18] is that we address delay bound violation probability, which makes our work more suitable for QoS provisioning for delay sensitive applications. In [19], an effective capacity approach was proposed to analyze the relationship

among data rate, delay bound and delay bound violation probability. However, the queueing model in [19] assumes infinite buffer space; and the large deviation method is used to derive the delay bound violation probability, and thus the results in [19] are only proved to hold in large delay regime while our results hold for arbitrary delay bound.

The remainder of the paper is organized as below. Section II describes the system model. In Section III, we present our analysis of packet error probability. Section IV presents the throughput maximization problem. Section V describes our RED theory. Simulation and numerical results are given in Section VI. Section VII concludes the paper.

II. System Model

We consider a joint queueing/coding single-user system as depicted in Fig. 1. Data packets, whose size is assumed to be L bits, arrive from some upper layer and are buffered at the link layer. Time is divided into blocks of equal length T_b. A rate control unit removes some head-of-line (HOL) packets from the buffer and convey them to the rate adaptive channel encoder in the physical layer. Then the encoded data is modulated and transmitted through a fading channel channel.

To elaborate, the channel at the physical layer is modeled as a discrete-time block-fading channel with additive white Gaussian noise. The transmitted signal is multiplied by a time-varying channel gain which models the fading. In each block, the channel gain is assumed to be fixed. We assume the (baseband) complex channel gain process $\{g_n\}$ is a stationary ergodic finite state Markov chain (FSMC) with state space $\mathcal{G}$. Such a channel model is suitable for modeling slowly varying, flat-fading channels [20] and is also adopted in [11] and [16]. Let W denote the bandwidth, then $N = WT_b$ modulated symbols can be transmitted in one block. Let $\mathbf{x} = (x_{n1}, x_{n2}, \ldots, x_{nN})$ denote the output of the rate adaptive channel encoder (coded symbols), which is the input of the modulator. Let $\mathbf{y} = (y_{n1}, y_{n2}, \ldots, y_{nN})$ denote the output of the demodulator in the n-th block. Then we have

$$y_{nk} = g_n x_{nk} + z_{nk}, \quad k = 1, 2, \ldots, N \tag{1}$$

where z_{nk} $(k = 1, 2, \ldots, N)$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variables with zero mean and variance N_0.

Note that we do not specify what channel encoding/decoding schemes are used[1], since our objective is to build a general framework for studying delay-constrained communication problems. We assume the codeword length does not exceed N (which is the number of channel uses per block), and a codeword needs to be decoded within one block. Thus the encoding/decoding delay should be no more than 2 blocks.

The queueing subsystem is modeled as a discrete-time finite-buffer queue with buffer size M packets. When newly arrived packets find that the buffer is full, some packets need to be dropped. There are three strategies for packet dropping: Strategy 1 (tail-dropping) drops the newly arrived packets; Strategy 2 (tail pushout) drops the tail (end-of-line) packets in the queue and appends the incoming packets to the tail of

[1]Different encoding/decoding schemes may result in different delay and error performances.

the queue; and Strategy 3 (HOL pushout) drops the head-of-line packets in the queue and appends the incoming packets to the tail of the queue. In the remainder of this paper, we always assume Strategy 3 (HOL pushout) is used.

Let $\{a_n\}$ be an i.i.d. random process with state space $\mathcal{A} \subset \mathbb{R}^+$ which represents the number of packets arriving at the buffer during the n-th block. We assume $\mathbb{E}[a_n] = \mu$ and, for the case of constant arrival, $a_n = \mu$. As shown in Fig. 2, at the very beginning of the n-th block, a batch of a_n packets arrive, followed by the departure of r_n (which we refer to as service rate or transmission rate) packets; the queue length q_n is observed immediately after the departure. Let $\mathcal{Q}$ denote the buffer state space, i.e., $\mathcal{Q} = \{0, 1, \ldots, M\}$.

In order to cope with channel variation, we assume the transmission rate is adaptive while the transmission power is constant. We assume the channel state information, buffer state information and arrival state information are available at both the transmitter and the receiver. Thus the transmission rate (or service rate) r_n is specified by a rate control policy $R : \mathcal{Q} \times \mathcal{G} \times \mathcal{A} \to \mathcal{Q}$, i.e.,

$$r_n = R(q_{n-1}, g_n, a_n). \tag{2}$$

Note that r_n depends on q_{n-1} instead of q_n because q_n is not available when the n-th departure takes place. So the evolution of the queueing system is given by

$$q_n = \min(q_{n-1} + a_n, M) - R(q_{n-1}, g_n, a_n) \tag{3}$$

where it is required that $R(q_{n-1}, g_n, a_n) \leq \min(q_{n-1} + a_n, M)$ (since R is a function of q_{n-1}, this requirement is natural). Let $\mathbf{s_n} \triangleq (q_{n-1}, g_n, a_n)$ denote the system state; then it is easy to see $\{\mathbf{s_n}\}$ forms a multivariate Markov chain.

Recall that we are interested in delay-constrained communication over fading channels. Let $D_{\max}$ denote the maximum tolerable delay, i.e., if one packet cannot reach its destination within $D_{\max}$ blocks, it will be considered as an erroneous packet. Let D denote the total delay experienced by a packet in the system in Fig. 1. Then D is the sum of the delay in the buffer plus the encoding/decoding delay. As we mentioned above, the encoding/decoding delay is confined to be at most 2 blocks. Thus we will omit the encoding/decoding delay but focus on the queueing delay in the remainder of this paper.

Our objective is to find the maximum system throughput while satisfying the delay and packet error probability constraints, which is equivalent to minimizing the packet error probability under the constraints on data rate (average arrival rate) and delay bound. Now consider a packet enters the system in Fig. 1. It may experience three types of errors: 1) packet drop due to full buffer, 2) delay bound violation (failing to reach the destination within $D_{\max}$ blocks), and 3) packet decoding error due to channel noise. It is easy to see that there is a tradeoff between the decoding error and the other two types of errors. If we increase (resp., decrease) the service rate, the buffer will be cleared more quickly (resp., slowly), resulting in a smaller (resp., larger) drop probability and delay bound violation probability; however, the decoding error probability will increase (resp., decrease) since more (resp., less) bits are transmitted through the channel. So the optimal rate control policy should balance packet drop error probability, decoding error

probability and delay bound violation probability so as to minimize the total error probability. In the next section, we analyze the probability of the three types of errors.

III. Analysis of Packet Error Probability

In this section, we first analyze each of the three types of errors then we give the expression of the total packet error probability.

A. Decoding Error Probability

Since the decoding error probability depends on channel gain g and service rate r (in unit of packets per block), we represent the decoding error probability as a function of g and r, denoted by $P_e^c(g,r)$. The specific expression of $P_e^c(g,r)$ depends on the modulation and channel coding schemes used [9], [21]. Next, we consider both block codes and convolutional codes.

For block codes, we only consider random coding and decoding. As we mentioned before, a codeword should be decoded at the end of a block. If r packets are to be transmitted in one block, then the service rate is rL/N bits per channel use, where L and N denote the packet size (in bits) and the number of channel uses per block, respectively. Assume the transmission energy is P_t per symbol which is constant for all symbols. We encode all bits to be transmitted in one block, into one single codeword. Then the following random coding bound [22] on the probability of decoding error holds, for any $\rho \in (0,1]$:

$$P_s \leq \exp(N(\rho rL/N \log 2 - E_0(\rho, g))) \tag{4}$$

where

$$E_0(\rho, g) = \rho \ln \left(1 + \frac{P_t |g|^2}{\sigma^2(1+\rho)}\right). \tag{5}$$

Since all bits to be transmitted in one block are encoded into one single codeword, if a codeword is decoded correctly, then there is no decoding error; otherwise, all information bits are un-decodable and erroneous. So we have for any $\rho \in (0,1]$:

$$P_e^c(g,r) \leq \exp(\rho rL \log 2 - NE_0(\rho, g)). \tag{6}$$

For convolution codes, the decoding error probability depends on the specified modulation/coding scheme. Assume BPSK is used (our analysis can be easily extended to any other linear modulation scheme). Let P_{bc} denote the bit error probability. It is bounded by [21]

$$P_{bc} \leq \sum_{d=d_{free}}^{\infty} B_d P_d \tag{7}$$

where B_d is the total number of nonzero information bits on all weight-d paths divided by the number of information bits per unit time; d_{free} is the minimal Hamming distance between different encoded

sequences; P_d is the pairwise error probability which is defined as the probability that the decoder selects an erroneous path at the distance d from the transmitted path. P_d depends on the channel type, modulation scheme and decoding type (hard or soft decision). For additive white Gaussian channel and hard decision decoding, the pairwise error probability is [21]

$$P_d = \begin{cases} \sum_{e=(d+1)/2}^{d} \binom{d}{e} (p_b)^e (1-p_b)^{d-e}, & d \text{ odd} \\ \frac{1}{2}\binom{d}{d/2}(p_b)^{d/2}(1-p_b)^{d/2} + \sum_{e=d/2+1}^{d} \binom{d}{e}(p_b)^e(1-p_b)^{d-e}, & d \text{ even}, \end{cases} \tag{8}$$

where p_b is the channel bit error rate; for BPSK, p_b is given by $p_b = Q(\sqrt{\frac{2E_b}{N_0}})$, where E_b is the energy per bit. Then we get an upper bound on packet decoding error probability

$$P_e^c(g, r) \leq 1 - (1 - P_{bc})^L. \tag{9}$$

In order to utilize the above upper bound, we need to choose appropriate convolution code which has required code rate. For BPSK modulation, the number of transmitted data bits is N in one block. The number of information bits to be transmitted is $L \times r$. So the required code rate is rL/N.

We define *average decoding error probability* as the ratio of the long-term average number of incorrectly decoded packets to the average number of arriving packets. Then the average decoding error probability is given by

$$\bar{P}_e^c = \limsup_{T\to\infty} \frac{1}{T\mu} \sum_{n=1}^{T} \mathbb{E}[P_e^c(g_n, r_n) r_n]. \tag{10}$$

Remark 1: Here, we only present the results for random coding and basic convolutional coding; the error analysis for other types of modulation/coding can be found in the literature.

B. Packet Drop Probability

When a packet arrives at the buffer, if the buffer is already full, the head-of-line packet will be dropped. We define *average packet drop probability* as the ratio of the long-term average number of dropped packets to the average number of arriving packets. In the n-th block, if the arrival rate and the previous queue length satisfy $q_{n-1} + a_n \geq M$, then $q_{n-1} + a_n - M$ packets will be dropped. So the average packet drop probability can be calculated by

$$\bar{P}_e^q = \limsup_{T\to\infty} \frac{1}{T\mu} \sum_{n=1}^{T} \mathbb{E}(q_{n-1} + a_n - M)^+ \tag{11}$$

where $(x)^+ \triangleq \max(x, 0)$. Note that the average packet drop probability depends on rate control policy which governs the queue state.

Remark 2: If, by choosing some rate control policy, the steady state of $\{\mathbf{s_n}\}$ exists, the packet drop probability is given by

$$\bar{P}_e^q = \frac{1}{\mu} \lim_{n\to\infty} \mathbb{E}(q_{n-1} + a_n - M)^+ \tag{12}$$

which can also be directly calculated by the standard FSMC approach, i.e., first calculating the steady state distribution of $\{\mathbf{s_n}\}$, then obtaining the packet drop probability via (12).

C. Delay Bound Violation Probability

We are interested in the delay experienced by a packet. The total delay D is the sum of the delay in the queue buffer and the encoding/decoding delay. As discussed in Section II, the encoding/decoding delay is at most 2 blocks, thus we can ignore the encoding/decoding delay and focus on the queueing delay. The queueing delay is the time that a packet spends in the buffer before it leaves the buffer. For example, if a packet arrives in the the m-th block and departs in the n-th block, then the queueing delay of the packet is $n-m$. The delay bound violation probability, or the deadline violation probability [23], is defined as the probability that a packet fails to reach its destination within a given delay bound. Generally, it is difficult to derive the queueing delay except for some simple queueing systems [24]. By analyzing the arrival and departure processes, numerical methods are proposed to calculate the queueing delay in [23] and [25]. Since we are considering a finite-buffer discrete-time queueing system, the departure, queue length and arrival of which are correlated, the calculation of queueing delay is, if not impossible, a formidable task. Thus we propose an alternative upper bound approach. Since the buffer size is finite, a packet can be either transmitted or dropped due to full buffer. Since the packet drop has been addressed in Section III-B, now we only consider the transmitted packets. We have the following useful lemma.

Lemma 1: Consider a discrete-time queueing system with arrival process a_n, departure process r_n, and the timing diagram shown in Fig. 2. If D denotes the queueing delay of a packet which departs in the n-th block, then the following inequalities hold.

$$\Pr\left(\sum_{i=n-D_{\max}+1}^{n} a_i < q_n\right) \le \Pr(D > D_{\max}) \le \Pr\left(\sum_{i=n-D_{\max}+1}^{n} a_i < q_n + r_n\right), \tag{13}$$

where q_n is the queue length in the n-th block and $D_{\max}$ is a positive delay bound.

Proof: We consider a packet τ which leaves the buffer in the n-th block. Note that $q_n + r_n$ is the queue length just before the n-th departure. The event $D > D_{\max}$ is equivalent to the event that packet τ arrives before the $(n - D_{\max} + 1)$-th block, which implies that the packets arriving in the i-th block $(i = n-D_{\max}+1, \ldots, n)$ are still in the buffer just before the n-th departure, i.e., $\sum_{i=n-D_{\max}+1}^{n} a_i < q_n + r_n$. The buffer status just before the n-th departure, is shown in Fig. 3. Thus we have

$$\Pr(D > D_{\max}) \le \Pr\left(\sum_{i=n-D_{\max}+1}^{n} a_i < q_n + r_n\right).$$

Now we prove the first inequality in (13). The inequality $\sum_{i=n-D_{\max}+1}^{n} a_i < q_n$ means that none of the packets arriving in the i-th block ($i = n - D_{\max} + 1, \ldots, n$) has left the buffer by the end of the n-th block. So packet τ, which leaves the buffer in the n-th block, must have arrived in a block prior to the $(n - D_{\max} + 1)$-th block (see Fig. 3), which implies $D > D_{\max}$. This proves the first inequality in (13). ■

Since $q_n + r_n \leq M$ and $\{a_i\}$ is stationary, from the second inequality in (13), we have the following upper bound

$$\Pr(D > D_{\max}) \leq \Pr\left(\sum_{i=1}^{D_{\max}} a_i < M\right). \tag{14}$$

Next, we consider three types of arrival processes, i.e., constant arrival, Poisson arrival and general i.i.d. arrival, and derive upper bounds on the delay bound violation probability.

a) Constant arrival: In this case, arrival rate a_i is a constant denoted by μ. Then (14) becomes

$$\Pr\left(D > D_{\max}\right) \leq \Pr\left(\mu \times D_{\max} < M\right) = \begin{cases} 0, & \text{if } D_{\max} \geq M/\mu, \\ 1, & \text{if } D_{\max} < M/\mu. \end{cases} \tag{15}$$

Thus for constant arrival, by choosing (virtual) buffer size M such that $M \leq \mu \times D_{\max}$, we can obtain zero delay bound violation probability, i.e., $\Pr\left(D > D_{\max}\right) = 0$.

b) Poisson arrival: We assume the arrival is a Poisson process with parameter μ, i.e., for all i,

$$\Pr(a_i = k) = \frac{e^{-\mu}\mu^k}{k!}, \quad k = 0, 1, \cdots$$

Since $\{a_i\}$ are i.i.d., $\sum_{i=1}^{D_{\max}} a_i$ is also Poisson distributed with parameter $\mu \times D_{\max}$. Then we have

$$\Pr\left(D > D_{\max}\right) \leq \sum_{i=0}^{M-1} \frac{e^{-D_{\max}\mu} D_{\max}^i \mu^i}{i!}. \tag{16}$$

c) General i.i.d. arrival: We assume $\{a_i\}$ are i.i.d. with mean μ and variance σ^2. Then by the well-known central limit theorem, as $D_{\max}$ approaches infinity, $1/(\sigma\sqrt{D_{\max}}) \sum_{i=1}^{D_{\max}} (a_i - \mu)$ converges in distribution to a Gaussian random variable of zero mean and unit variance. Thus we have the following approximation for large delay bound $D_{\max}$

$$\Pr\left(D > D_{\max}\right) \leq \Pr\left(\sum_{i=1}^{D_{\max}} a_i < M\right) \approx 1 - Q\left(\frac{\sqrt{D_{\max}}}{\sigma}\left(\frac{M}{D_{\max}} - \mu\right)\right), \tag{17}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{\frac{-t^2}{2}} dt$.

Remark 3: It is easy to see that the delay bound violation probability is non-increasing in the delay bound. Given delay bound $D_{\max}$, by choosing a larger buffer size M, the upper bound on delay bound

violation probability is increased while the packet drop probability is decreased. Thus there is a tradeoff between delay bound violation probability and packet drop probability. This interesting tradeoff, with different model and assumptions, has been studied in [23].

Remark 4: In this section, we have derived the upper bound on delay bound violation probability for i.i.d. arrival process. In fact, if the arrivals are correlated, e.g., $\{a_n\}$ is a Markov chain, we can also calculate the corresponding upper bound based on large deviation theory [26]. An upper bound on $\Pr\left(\sum_{i=1}^{D_{\max}} a_i \leq M\right)$ can be derived by directly applying the Gartner-Ellis theorem.

Remark 5: In simulations, we observe that our upper bound on delay bound violation probability is tighter when the load is higher (i.e., the average arrival rate is closer to the ergodic channel capacity); and the upper bound is looser when the load is lower. In addition, we also observe that the value of our upper bound becomes very small if the (virtual) buffer size M is chosen to be smaller than $\mu \times D_{\max}$. Thus, we can choose appropriate buffer size so that the delay bound violation probability is negligible, which simplifies the analysis in the remainder of this paper.

We define *average delay bound violation probability* as the ratio of the long-term average number of packets that violate delay bound, to the average number of arriving packets. Then the average delay bound violation probability can be calculated by

$$\bar{P}_e^d = \limsup_{T\to\infty} \frac{1}{T\mu} \sum_{n=1}^{T} \mathbb{E}[\Pr\left(D > D_{\max}\right) r_n] \tag{18}$$

$$\leq \hat{P}_e^d \limsup_{T\to\infty} \frac{1}{T\mu} \sum_{n=1}^{T} \mathbb{E}[r_n] \tag{19}$$

$$= \hat{P}_e^d \times (1 - \bar{P}_e^q) \tag{20}$$

where $\hat{P}_e^d$ denotes the upper bound on $\Pr\left(D > D_{\max}\right)$, which we derived in this section, and the last equality holds because $\lim_{n\to\infty} \mathbb{E}[r_n] = \mu \times (1 - \bar{P}_e^q)$, that is, the transmission rate is equal to the arrival rate minus the drop rate.

Remark 6: The delay bound violation probability may be different for different blocks. However, the upper bound on delay bound violation probability derived in this section holds for any block, which guarantees the correctness of (19).

Remark 7: In the above discussion, we implicitly assumed that if a packet violations its delay bound, it will still be transmitted. However, in practical systems, a better choice may be for transmitter to drop a packet that violates its delay bound since the packet has already been regarded as an erroneous packet. Even under this scenario, the upper bounds on decoding error probability, packet drop probability and delay bound violation probability still hold; so do the results in the remainder of this paper.

D. Total Packet Error Probability

Based on the above error analysis, now we are ready to calculate the total packet error probability (or simply *packet error probability*), which is defined as the ratio of the long-term average number of erroneous packets to the long-term average number of arriving packets. Now consider a packet τ, we have

$$\Pr(\tau \text{ experiences packet error}) = \Pr(\tau \text{ is dropped or incorrectly decoded or violates delay bound}). \tag{21}$$

Let $P_e(R, D_{\max}, \mu)$ denote the total packet error probability, which is a function of rate control policy R, delay bound $D_{\max}$ and average arrival rate μ. Then we have the following union bound

$$P_e(R, D_{\max}, \mu) \leq \bar{P}_e^q + \bar{P}_e^d + \bar{P}_e^c \tag{22}$$

where $\bar{P}_e^c, \bar{P}_e^q$ and $\bar{P}_e^d$ are given by (10), (11) and (20), respectively. Especially, if the arrival is constant and the (virtual) buffer size M is chosen to be $\mu \times D_{\max}$, we have

$$P_e(R, D_{\max}, \mu) = \bar{P}_e^q + \bar{P}_e^c, \tag{23}$$

where the equality holds because packet drop and decoding error are mutual exclusive, i.e., if one packet is dropped, it cannot be transmitted.

In this section, we have conducted packet error analysis. The total packet error probability is expressed as a function of rate control policy, delay bound and average arrival rate. In the next section, we will investigate the optimal rate control policy that minimizes the total packet error probability.

IV. Throughput Maximization Problem

As in [16], the throughput of the system in Fig. 1 is defined as the long-term average data rate at which packets are successfully transmitted. Given a random arrival process $\{a_n\}$ with mean μ and delay bound $D_{\max}$, the error-free throughput can be calculated by

$$\mu \times (1 - P_e(R, D_{\max}, \mu)). \tag{24}$$

Thus, maximizing throughput is equivalent to minimizing packet error probability, i.e.,

$$\min_{R \in \mathcal{R}} P_e(R, D_{\max}, \mu) \tag{25}$$

where $\mathcal{R}$ is the space of admissible rate control policies. Since the exact packet error probability is unavailable, we seek to find a rate control policy which minimizes the following upper bound

$$\limsup_{T \to \infty} \frac{1}{T\mu} \sum_{n=1}^{T} \mathbb{E}\left(P_e^c(g_n, r_n) r_n + (1 - \hat{P}_e^d)(q_{n-1} + a_n - M)^+ + \mu \hat{P}_e^d \right), \tag{26}$$

which is derived based on (22). From the above discussion, the problem of finding the rate control policy that maximizes the error-free throughput is an average cost Markov decision problem with state space

$\mathcal{Q}\times\mathcal{G}\times\mathcal{A}$ and per-stage cost $(P_e^c(g_n,r_n)r_n+(1-\hat{P}_e^d)(q_{n-1}+a_n-M)^+ +\mu\hat{P}_e^d)/\mu$. Such an optimization problem can be solved by the policy iteration algorithm [27]. However, the computational complexity of policy iteration is high. The computational complexity is $O(|\mathcal{Q}|^3|\mathcal{G}|^3|\mathcal{A}|^3)$. Fortunately, in our simulations, we find the policy iteration algorithm converges in a small number of iterations, e.g., in about 10 iterations. Thus the computational burden is acceptable usually.

V. RED Triplet and Its Properties

In this section, we study the throughput maximization problem from another perspective. We want to find the maximum constant rate that can be supported by a time-varying channel under delay and packet error constraints, which is similar to the problems studied by the effective bandwidth [28] and effective capacity [19] approaches. This problem can also be interpreted as identifying the maximum rate of a constant-rate equivalent pipe/channel, which achieves the same delay and error performances as that achieved by a given time-varying channel[2]. We propose to use a triplet $(\mu, D_{\max}, \varepsilon)$ to characterize the delay-constrained throughput of a fading channel, where μ is the maximum data rate of a flow with delay bound $D_{\max}$ and packet error probability ε. Specifically, we want to find the maximum data rate μ achievable under delay bound $D_{\max}$ and error probability ε constraints. By varying the delay and packet error probability constraints, we obtain a Pareto optimal surface. Such a rate-error-delay (RED) triplet completely describes the performance of delay-constrained communication over fading channels.

As we mentioned previously, if the arrival is constant and the virtual buffer size $M \le \mu \times D_{\max}$, then $\Pr(D > D_{\max}) = 0$. Since the virtual buffer size[3] M can take any positive integer value as desired, we always choose $M = \mu \times D_{\max}$. In the remainder of this paper, we always make the following two assumptions: 1) constant arrival, 2) buffer size $M = D_{\max}\times\mu$. Thus, the rate control policy R only depends on the channel gain g_n and queue length q_{n-1}, i.e., the departure rate in the n-th block is $r_n = R(q_{n-1}, g_n)$.

Next, we study the structure of the admissible control space of the rate control policy in Section V-A; and give some properties of the RED surface in Section V-B.

A. Admissible Control Space

So far, the rate control policy R is assumed to be stationary. In general, the rate control policy can be time-varying. A general rate control policy can be denoted by $\mathbf{u} = \{u_1, u_2, \ldots\}$, that is, the departure rate in the n-th interval is $r_n = u_n(q_{n-1}, g_n)$. The evolution of the queueing system becomes

$$q_n = \min(q_{n-1} + \mu, M) - u_n(q_{n-1}, g_n) \tag{27}$$

[2]When a user watches streaming video, he/she does not care whether the physical layer channel is wireless or coaxial cable or twisted pair or Ethernet cable. Hence, we can convert any time varying channel into a a constant-rate equivalent pipe/channel, and study the QoS achieved by this constant-rate equivalent pipe to simplify the analysis. This is exactly the idea behind the effective bandwidth and the effective capacity approaches.

[3]A virtual buffer is not a physical buffer. A virtual buffer can be implemented by a linked list with its maximum length equal to virtual buffer size M, where M can be any positive integer chosen by the user.

Let $U(q_{n-1}, g_n) \subseteq \mathcal{Q}$ denote the set of feasible values of u_n, i.e., $u_n(q_{n-1}, g_n) \in U(q_{n-1}, g_n)$. Now we determine $U(q_{n-1}, g_n)$. Since the queue length is available at the transmitter, the departure rate (in unit of packets/block) cannot exceed the total number of packets in the buffer, so we have $u_n(q_{n-1}, g_n) \leq \min(q_{n-1} + \mu, M)$. Moreover, since the CSI is available at the transmitter, it is not reasonable to transmit at a rate larger than the instantaneous channel capacity. So we have $u_n(q_{n-1}, g_n) \leq C(g_n)$, where

$$C(g_n) = \lfloor \frac{N}{L} \log_2 \left(1 + |g_n|^2 \times \frac{P_t}{N_0 W}\right) \rfloor. \tag{28}$$

Thus we have

$$\begin{aligned} U(q_{n-1}, g_n) &= \{r | r \in \mathcal{M}, r \leq \min(q_{n-1} + \mu, M), r \leq C(g_n)\} \\ &= \{0, 1, \cdots, \min(M, q_{n-1} + \mu, C(g_n))\} \end{aligned} \tag{29}$$

We denote by Π the set of admissible policies, which is the set of all sequences of functions $\mathbf{u} = \{u_1, u_2, \ldots\}$ where $u_n : \mathcal{Q} \times \mathcal{G} \to \mathcal{Q}$ and $u_n(q_{n-1}, g_n) \in U(q_{n-1}, g_n)$, $n = 1, 2, \ldots$ Then the minimum packet error probability is

$$\min_{\mathbf{u} \in \Pi} P_e(\mathbf{u}, D_{\max}, \mu). \tag{30}$$

As we discussed in Section IV, the above maximization problem is an infinite-horizon Markov decision problem with average cost. From [27], it is known that for an average cost problem with finite state and control space, there always exists an optimal stationary policy. For infinite state and control space, with some mild conditions, there also exists an optimal stationary policy. Without loss of generality, throughout this paper, we assume there always exists an optimal stationary rate control policy.

Now we present two lemmas about the structure of the admissible control space Π. Consider a system as shown in Fig. 1, given a realization of channel gain sequence $\mathbf{H} = \{H_n\}$ and buffer size M (note that we use different symbols to denote the random variables and their realizations), we say a sequence of control actions $\{r_n\}$ is feasible if $r_n \in U(q_{n-1}, H_n)$, where $\{q_n\}$ is the corresponding queue length sequence. The feasible condition guarantees $q_n \geq 0, n = 0, 1, \ldots$ Let $\Gamma_M^{\mathbf{H}}$ denote the set of feasible control action sequences. We have the following results.

Lemma 2: Given a realization of channel gain sequence $\mathbf{H}$ and two buffer sizes M^1, M^2. If $M^1 < M^2$, then $\Gamma_{M^1}^{\mathbf{H}} \subset \Gamma_{M^2}^{\mathbf{H}}$.

For a proof, see the Appendix. Lemma 2 tells that with a larger buffer size, we will have more freedom to choose control policies. Now we study the extreme case, that is, infinite buffer size. In this case, the buffer size constraint on the feasible control space is removed. We have the following result for infinite buffer systems.

Lemma 3: Given a realization of channel gain sequence $\mathbf{H}$, consider the system shown in Fig. 1 with infinite buffer size. The set of feasible control action sequences $\Gamma_\infty^{\mathbf{H}}$ is convex.

For a proof, see the Appendix. If the buffer size is finite, Lemma 3 may not hold.

B. ERD function and RED function

In order to study the RED Pareto surface, we define two functions: ERD and RED. The ERD function represents the minimum achievable packet error probability under rate and delay constraints. Specifically, the ERD function represents the minimum achievable packet error probability as a function of data rate μ and delay bound $D_{\max}$, i.e.,

$$ERD(\mu, D_{\max}) \triangleq \min_{R \in \mathcal{R}} P_e(R, D_{\max}, \mu). \tag{31}$$

The RED function represents the maximum achievable data rate under delay and packet error probability constraints. Specifically, the RED function represents the maximum achievable data rate as a function of packet error probability ε and delay bound $D_{\max}$, i.e.,

$$\begin{aligned} RED(\varepsilon, D_{\max}) \quad &\triangleq \quad \max \ \mu \\ &\qquad \text{subject to } ERD(\mu, D_{\max}) \leq \varepsilon. \end{aligned} \tag{32}$$

Next we investigate the properties of ERD and RED functions. We have the following proposition about monotonicity of the ERD function.

Proposition 1: The ERD function is a monotonically decreasing function of delay bound $D_{\max}$ for fixed μ.

For a proof, see the Appendix.

Now assume the buffer size is infinite, i.e., the queue can grow to infinite. For infinite buffer, the ERD function has nicer properties as below.

Proposition 2: If the buffer size is infinite, the ERD function is a monotonically decreasing function of delay bound $D_{\max}$ for fixed μ. If the decoding error probability $P_e^c(g, r)$ is a convex function of rate r, the ERD function is also convex in $D_{\max}$ for fixed μ.

For a proof, see the Appendix.

In Proposition 2, the requirement that $P_e^c(g, r)$ is a convex function of r can be satisfied with random coding, which can be seen from (6). Proposition 2 tells that although the minimum packet error probability is not a convex function of delay bound in general, an upper bound on the minimum packet error probability is convex.

The following theorem presents the monotonic property of the RED function and the RED Pareto surface.

Proposition 3: The RED function $RED(\varepsilon, D_{\max})$ is an monotonically increasing function of $D_{\max}$ for fixed ε, and an monotonically increasing function of ε for fixed $D_{\max}$.

For a proof, see the Appendix.

Remark 8: It is easy to see that, given an RED triplet $(\mu, D_{\max}, \varepsilon)$, the maximum error-free throughput is $\mu \times (1 - \varepsilon)$.

VI. Simulation and Numerical Results

In this section, simulation and numerical results are presented to illustrate our theoretical results. After giving the experimental settings in Section VI-A, we use simulation results to verify the upper bound on delay bound violation probability proposed in Section III-C. Then we use an example to show the tradeoff between packet drop probability (link layer) and decoding error probability (physical layer), which justifies a major means in our RED theory, i.e., an optimal rate control policy that optimally balances the link layer and the physical layer performance. Then, we give an example of the Pareto surface, resulting from the optimal rate control policy. Finally, we compare the optimal policy with the optimal fixed-decoding-error policy.

A. Simulation settings

We simulate the joint queueing/coding system as depicted in Fig. 1. We consider two types of arrival: constant arrival and Poisson arrival, both of which have mean μ packets per block. The departure is determined by a rate control policy R. The channel gain sequence $\{g_n\}$ is modeled as a Markov chain with state transition matrix Q. In all our experiments, the time block length is set to $T_b = 0.005$ s. The channel bandwidth is $W = 20$ kHz, thus the number of complex channel uses per block $N = WT_b = 100$. The SNR without fading (channel gain $|g_n| = 1$) is 10 dB. The link layer packet size is set to 10 bits. The set of simulation setting can be arbitrary, however, we just choose appropriate parameters to clearly show our results.

We consider three types of rate control policies: linear policy, optimal policy, and optimal fixed-decoding-error policy. The linear rate control policy is one of the simplest policies and is defined as below

$$R_{Linear}(m, g) = \min(m + \mu, M, \lfloor \rho \times C(g) \rfloor) \tag{33}$$

where $C(g)$ is the instantaneous channel capacity, and $\rho \in [0, 1]$. We call it linear policy since the rate is upper bounded by a linear function of the instantaneous channel capacity, i.e., $\rho \times C(g)$.

The optimal policy can be obtained by the policy iteration algorithm [27]. The optimal fixed-decoding-error policy is an optimal policy under the constraint of fixed decoding error probability. For a target decoding error probability $\epsilon_{dec} \in (0, 1)$, the optimal fixed-decoding-error policy is defined by

$$R_{DEC}(m, g) = \min\left(m + \mu, M, \lfloor \max_{P_e^c(g,r) \le \epsilon_{dec}} r \rfloor\right). \tag{34}$$

The optimal fixed-decoding-error policy takes a "pure queueing approach" since it only optimizes the queueing performance while keeping fixed decoding error probability; in other words, it does not jointly optimize the physical layer and the link layer.

B. Delay Bound Violation Probability

In this section, we simulate the queueing subsystem. The arrival is assumed to Poisson distribution with mean $\mu = 12$ or 10 packets/block; buffer size $M = 100$ packets; the channel gain process has two states $G_1 = 1$, $G_2 = 0.3$ and transition probability matrix is

$$Q = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}. \tag{35}$$

The rate control policy is an linear policy with parameter $\rho = 0.4$ as defined in (33). Fig. 4 shows the simulation result of delay bound violation probability compared with the upper bound proposed in (16) in Section III-C.

From the figure, we can see that both the simulation result and the upper bound decrease rapidly as the delay bound becomes larger, especially when the delay bound is larger than M/μ. Thus we can choose an appropriate delay bound (e.g., $2 \times M/\mu$) such that we can omit the delay bound violation probability and thus simplify our analysis.

C. Tradeoff between Decoding Error Probability and Packet Drop Probability

In this simulation, the channel gain process has two states $G_1 = 1$, $G_2 = 0.5$ and state transition probability matrix is

$$Q = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}.$$

We use linear rate control policies defined in (33). Fig. 5 shows packet drop probability, average decoding error probability, and average packet error probability vs. ρ (the parameter of the linear control policy). From the figure, it can be observed that as ρ increases, packet drop probability decreases while decoding error probability increases. Hence, there is a tradeoff between packet drop probability and decoding error probability. In addition, the figure shows that the minimum packet error probability is achieved at $\rho = 0.7$; i.e., for this system, the optimal linear rate control policy that minimizes the packet error probability is

$$R^*_{Linear}(m, g) = \min(m + \mu, M, \lfloor 0.7 * C(g) \rfloor) \tag{36}$$

This example indicates that the optimal performance can only be achieved through cross layer design due to the conflicting nature between the link layer and the physical layer performance (i.e., packet drop probability vs. decoding error probability).

D. An Example of Pareto Surface

In this simulation, the channel gain and transition matrix are the same as those in Section III-C. It is easy to calculate that the Shannon ergodic capacity (without power control) is 44 kb/s; for rate control, we use the optimal policy. Fig. 6 shows the Pareto surface, i.e., maximum data rate μ as a function of delay bound $D_{\max}$ and packet error probability. In this and the next simulation, when plotting Pareto surfaces,

we increase the delay bound by 2 blocks to include the encoding/decoding delay. From this figure, it can be observed that the maximum data rate is a monotonically increasing function of delay bound and packet error probability. Moreover, the maximum data rate is always smaller than the ergodic capacity.

E. Optimal Policy vs. Optimal Fixed-decoding-error Policy

In this simulation, the channel gain and transition matrix are the same as those in the previous section; the optimal policy can be obtained by the policy iteration algorithm; the optimal fixed-decoding-error policy is defined in (34), and we choose $\epsilon_{dec} = 10^{-5}$. Fig. 7 shows the maximum data rate as a function of delay bound $D_{\max}$ and packet error probability, under the optimal policy and the optimal fixed-decoding-error policy, respectively. From this figure, it can be seen that the optimal policy achieves higher data rate than the optimal fixed-decoding-error policy. This is because the optimal policy balances packet drop probability (link layer) and decoding error probability (physical layer) while the optimal fixed-decoding-error policy is a pure queueing (link layer) approach and only minimizes packet drop probability. Again, it shows the superiority of cross layer design over optimizing each layer individually.

VII. Conclusion

In this work, we studied the problem of data communication with both delay and packet error probability constraints. The transmission data rate is adapted to channel state, buffer state and arrival state to minimize the total packet error probability thus maximize the system throughput. Different from most previous works, we considered a system with finite buffer space thus we addressed three types of errors: 1) packet drop due to full buffer, 2) delay bound violation, and 3) packet decoding error due to channel noise. We derived an upper bound on the total packet error probability. By minimizing the packet error probability over the transmission rate, we obtained an optimal rate control policy that guarantees the user-specified data rate and delay bound. Then by varying data rate and delay bound, we obtained RED Pareto-optimal surface. Our results provide important insights into statistical QoS provisioning in wireless systems; the RED Pareto surface represents a major step towards deriving the probabilistic delay-constrained channel capacity of fading channels. In our future work, both rate adaptation and power adaptation will be used to achieve a higher data rate.

Appendix

Proof of Lemma 2: Consider two systems, A and B, which have the same structure as shown in Fig. 1 but different buffer sizes. System A has a buffer of size M^1 and System B has a buffer of size M^2, where $M^1 < M^2$. Given a realization of channel gain sequence $\mathbf{H} = \{H_n\}$, to show that $\Gamma^{\mathbf{H}}_{M^1} \subset \Gamma^{\mathbf{H}}_{M^2}$, we just need to prove that if a sequence of control actions $\{r_n\}$ is feasible for System A, then it is also feasible for System B.

Let $\{q^i_n\},\ i = 1, 2$ denote the queue length sequences of System A and System B respectively. Since $\{r_n\}$ is feasible for System A, we have $r_n \in U(q^1_{n-1}, H_n)$, i.e.,

$$r_n \in [0, \min(M^1, q^1_{n-1} + \mu, C(H_n))],\ n = 1, 2, \ldots$$

We need to show that

$$r_n \in [0, \min(M^2, q_{n-1}^2 + \mu, C(H_n))],\ n = 1, 2, \ldots$$

Thus we just need to prove that $q_{n-1}^1 \leq q_{n-1}^2, n = 1, 2, \ldots$ We prove this by induction. Without loss of generality, we assume $q_0^1 = q_0^2 = 0$. Assume for some $k > 0$, $q_k^1 \leq q_k^2$, we have

$$\begin{aligned} q_{k+1}^2 &= \min(q_k^2 + \mu - r_{k+1}, M^2) \\ &\geq \min(q_k^1 + \mu - r_{k+1}, M^2) \\ &\geq \min(q_k^1 + \mu - r_{k+1}, M^1) = q_{k+1}^1 \end{aligned}$$

This completes the proof. ■

Proof of Lemma 3: Given a realization of channel gain sequence $\mathbf{H} = \{H_n\}$, consider the system shown in Fig. 1 with infinite buffer size. To show the set of feasible control action sequences $\Gamma_\infty^{\mathbf{H}}$ is a convex set, we just need to prove that if two sequences of control actions $\mathbf{u^i} = \{r_n^i\},\ i = 1, 2$ are feasible, then for any $\lambda \in [0, 1]$, $\mathbf{u}^\lambda = \lambda * \mathbf{u^1} + (1 - \lambda)\mathbf{u^2}$ is also feasible.

Let $q^i, i = 1, 2, \lambda$ denote the queue length sequences of an infinite-buffer queueing system with same arrival rate μ but different departure sequences $\{r_n^i\}, i = 1, 2, \lambda$. Then we have

$$q_n^i = q_{n-1}^i + \mu - r_n^i,\ i = 1, 2, \lambda.$$

By induction, it is easy to show that $q_n^\lambda = \lambda q_n^1 + (1 - \lambda)q_n^2$. Since $r_n^i \in [0, \min(q_{n-1}^i + \mu, C(H_n))],\ n = 1, 2, \ldots$ (in this case, the constraint on buffer size is lifted), we get $\lambda r_n^1 + (1 - \lambda)r_n^2 \in [0, \min(q_{n-1}^\lambda + \mu, C(H_n))]$, thus $\mathbf{u}^\lambda$ is also feasible. ■

Proof of Proposition 1: Let D_1, D_2 denote two delay bounds. Without loss of generality, we assume $D_1 < D_2$. Consider two systems: System A with buffer size M^1 and System B with buffer size M^2, where $M^i = \mu \times D_i, i = 1, 2$. We just need to show $ERD(D_1, \mu) \geq ERD(D_2, \mu)$. We prove this by sample path argument. Let $\mathbf{H} = \{H_n\}$ be a given realization of channel gain sequence. Assume the optimal rate control action sequence is $\mathbf{u} = \{r_1, r_2, \ldots\}$ for System A, i.e., $ERD(D_1, \mu) = P_e(\mathbf{u}, D_1, \mu)$. Since $M^2 > M^1$, from Lemma 2, we see that $\mathbf{u}$ is also an admissible control action sequence for System B. Since the average number of packet drop in one block equals to the arrival rate minus the average departure rate. Moreover, with same rate control policy, the number of packets incorrectly decoded is also the same, so we have

$$P_e(\mathbf{u}, D_1, \mu) = P_e(\mathbf{u}, D_2, \mu)$$

So we get

$$ERD(D_2, \mu) \leq P_e(\mathbf{u}, D_2, \mu) = ERD(D_1, \mu).$$

Thus we see that the ERD function is a monotonically decreasing function of D. ■

Proof of Proposition 2: We omit the proof to the first part since it is similar to Proposition 1.

Now, we prove the second part with sample path argument. Since the delay bound and buffer size is related by $M = \mu \times D_{\max}$, it is suffice to show that $ERD(M/\mu, \mu)$ is convex in M for fixed μ. Let $\mathbf{H} = \{H_n\}$ be a given realization of channel gain sequence. Consider an infinite-buffer system and two virtual buffer sizes $M^i, i = 1, 2$. Let $0 \leq \lambda \leq 1$ and $M^\lambda = \lambda M^1 + (1-\lambda)M^2$. We need to show that $ERD(M^\lambda/\mu, \mu) \leq \lambda ERD(M^1/\mu, \mu) + (1-\lambda)ERD(M^2/\mu, \mu)$.

Assume the control action sequences that attain $ERD(M^i/\mu, \mu), i = 1, 2$ are $\mathbf{u^i} = \{r_n^i\}, i = 1, 2$ and the corresponding queue length sequences are $\{q_n^i\}, i = 1, 2$. Now consider virtual buffer size M^λ. Let the control action sequence be $\{r_n^\lambda\}$, where $r_n^\lambda = \lambda r^1 + (1-\lambda)r^2$. From Lemma 3, we see that $\{r_n^\lambda\}$ is also feasible and the corresponding queue length sequence is $q_n^\lambda = \lambda q_n^1 + (1-\lambda)q_n^2$. Since $(\cdot)^+$ is a convex function, we have

$$(q_n^\lambda + \mu - M^\lambda)^+ \leq \lambda(q_n^1 + \mu - M^1)^+ + (1-\lambda)(q_n^2 + \mu - M^2)^+.$$

Since $P_e^c(g, r)$ is convex in r for fixed g, it is easy to show (by taking derivation on r) that $rP_e^c(g, r)$ is also convex in r for $r \geq 0$. So we have

$$r_n^\lambda P_e^c(H_n, r_n^\lambda) \leq \lambda r_n^1 P_e^c(H_n, r_n^1) + (1-\lambda) r_n^2 P_e^c(H_n, r_n^2).$$

From (23), we obtain

$$ERD(M^\lambda/\mu, \mu) \leq \lambda ERD(M^1/\mu, \mu) + (1-\lambda)ERD(M^2/\mu, \mu).$$

That is, for fixed arrival rate μ, $ERD(D_{\max}, \mu)$ is a convex function of $D_{\max}$. ■

Proof of Proposition 3: Let $D^i,\ i = 1, 2,\ D^1 < D^2$ be two delay constraints and $\varepsilon > 0$. Let $\mu^1 = RED(D^1, \varepsilon)$, we have $ERD(D^1, \mu^1) \leq \varepsilon$. From Proposition 1, we see

$$ERD(D^2, \mu^1) \leq ERD(D^1, \mu^1) \leq \varepsilon$$

So we obtain $RED(D^2, \varepsilon) \geq RED(D^1, \varepsilon)$. Similarly, we can prove $RED(D_{\max}, \varepsilon)$ is an increasing function of ε for fixed $D_{\max}$. ■

REFERENCES

[1] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[2] ——, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 868–886, May 1996.

[3] G. Caire and S. Shamai(Shitz), "On the capacity of some channels with channel state information," *IEEE Transactions on Information Theory*, vol. 45, pp. 2007–2019, 1999.

[4] A. J. Goldsmith and M. Mdard, "Capacity of time-varying channels with causal channel side information," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 881–899, 2007.

[5] H. Viswanathan, "Capacity of markov channels with receiver csi and delayed feedback," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 761–771, 1999.

[6] T. Yoo and A. J. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[7] A. J. Goldsmith and S.-G. Chua, "Adaptive coded modulation for fading channels," *IEEE Transactions on Communications*, vol. 46, no. 5, pp. 595–602, May 1998.

[8] ——, "Variable-rate variable-power MQAM for fading channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[9] B. Vucetic, "An adaptive coding scheme for time-varying channels," *IEEE Transactions on Communications*, vol. 39, no. 5, pp. 653–663, May 1991.

[10] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channelscpart II:delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2830, 1998.

[11] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[12] I. Bettesh and S. S. (Shitz), "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4115–4141, 2006.

[13] M. Goyal, A. Kumar, and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," *Proc. IEEE INFOCOM'03*, 2003.

[14] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal energy allocation for delay-constrained data transmission over a time-varying channel," *Proc. IEEE INFOCOM'03*, 2003.

[15] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks-part I: the fluid model," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2568–2592, 2006.

[16] A. T. Hoang and M. Motani, "Cross-layer adaptive transmission: Optimal strategies in fading channels," *IEEE Transactions on Communications*, vol. 56, no. 5, pp. 799–807, May 2008.

[17] Q. Liu, S. Zhou, and G. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, 2004.

[18] ——, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, 2005.

[19] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[20] H. Wang and N. Moayeri, "Finite-state Markov channel–a useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, 1995.

[21] S. Lin and D. J. Costello, *Error Control Coding*. Prentice Hall, 2004.

[22] R. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.

[23] M. Balakrishnan, A. Puliafito, K. Trivedi, and Y. Viniotis, "Buffer losses vs. deadline violations for ABR traffic in an ATM switch: A computational approach," *Telecommunication Systems*, vol. 7, no. 1, pp. 105–123, 1997.

[24] L. Kleinrock and R. Gail, *Queueing Systems*. Wiley New York, 1976.

[25] F. Albizuri, M. Graña, and B. Raducanu, "Statistical transmission delay guarantee for nonreal-time traffic multiplexed with real-time traffic," *Computer Communications*, vol. 26, no. 12, pp. 1365–1375, 2003.

[26] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer Verlag, 2009.

[27] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 1995, vol. 2.

[28] C.-S. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, Aug. 1995.

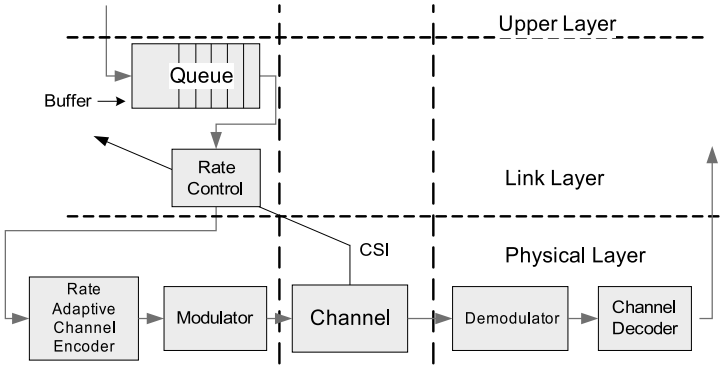

Fig. 1: System model

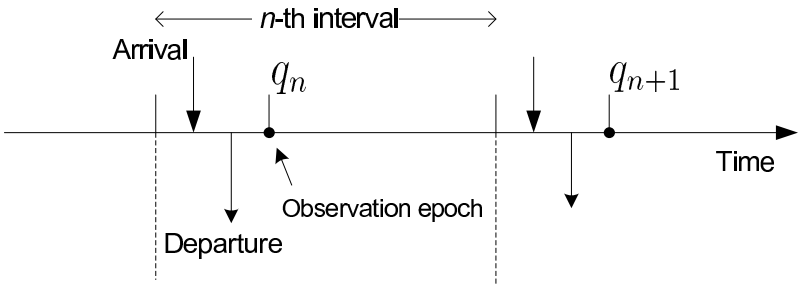

Fig. 2: Timing diagram

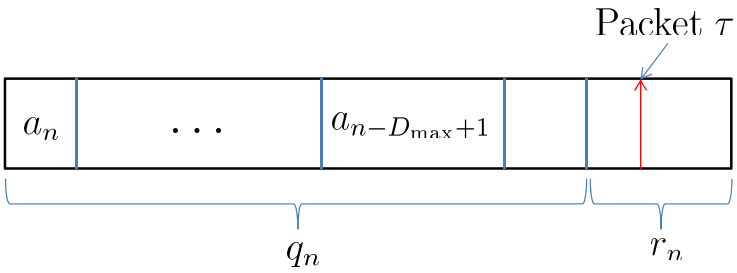

Fig. 3: Buffer status in the n-th block

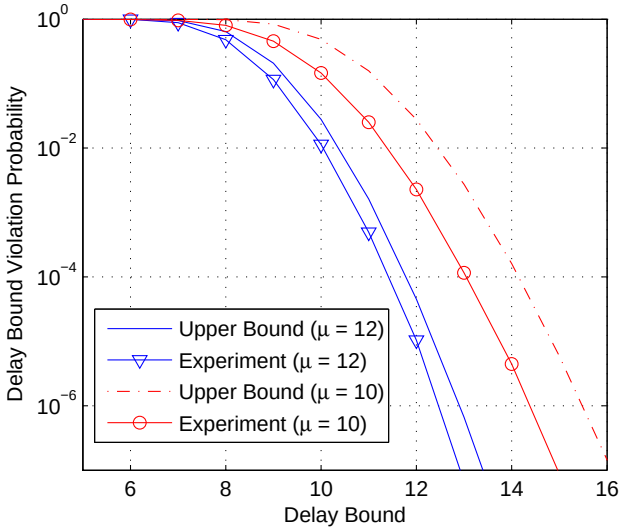

Fig. 4: Delay bound violation probability and its upper bound

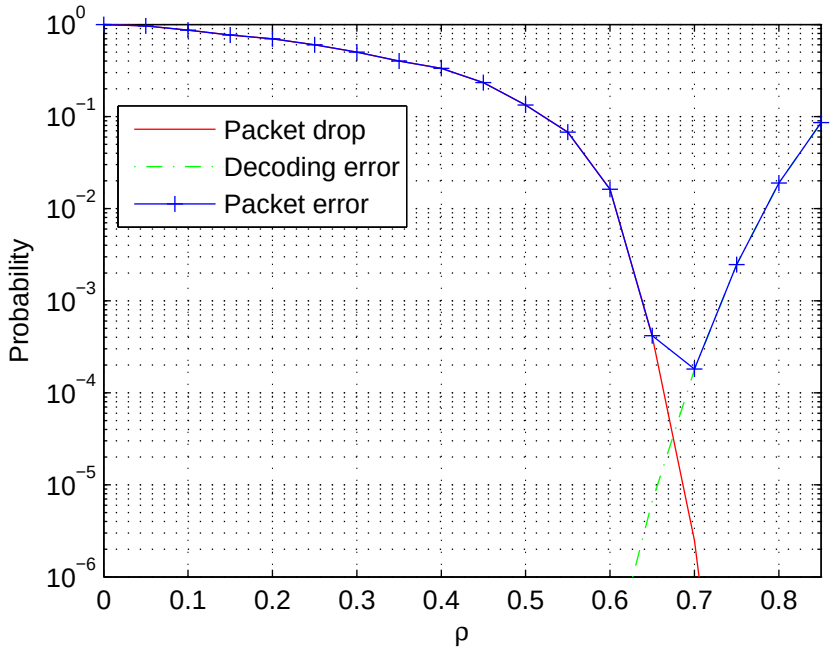

Fig. 5: Packet error probability, decoding error probability, and packet drop probability under linear rate control policies

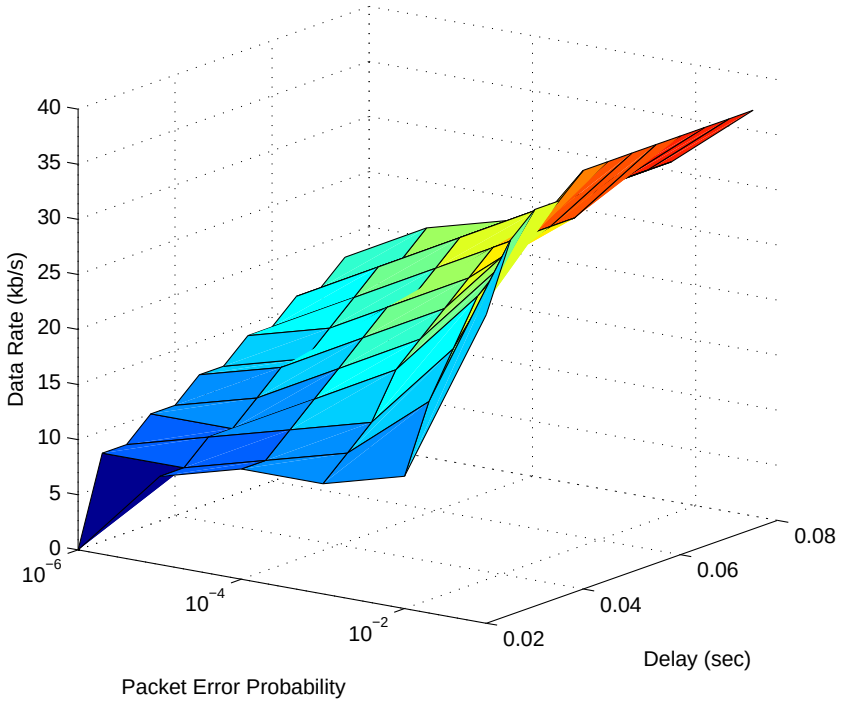

Fig. 6: Pareto surface: maximum data rate (μ) as a function of delay bound D and packet error probability ε

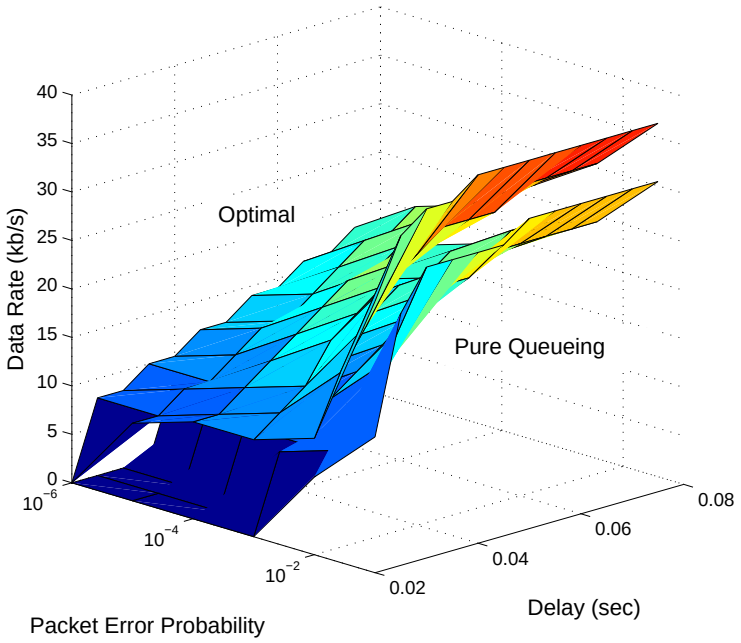

Fig. 7: Performance comparison between the optimal policy and the optimal fixed-decoding-error policy (pure queueing)