

An Information-Theoretic Approach to Queuing in Wireless Channels with Large Delay Bounds

Rohit Negi, Satashu Goel
negi@ece.cmu.edu, satashug@ece.cmu.edu
Department of Electrical and Computer Engineering
Carnegie Mellon University

Abstract—Queuing theory allows the design of communication links that can provision for Quality of Service (QoS) in time-varying channels, such as mobile wireless channels, by considering an idealized queuing system that abstracts out the physical layer. Similarly, information theory allows the design and analysis of channel codes that can guarantee low decoding error probability by adapting to the time-varying channel. Whereas other researchers have attempted cross-layer design methods that combine these two approaches, these have been limited to specific choices of practical channel codes. There have been few attempts to combine these two theories to specify the ultimate limit of delay-constrained communications. This paper presents an approach that obtains a QoS exponent, by combining the queuing and information theoretic models; in particular, by considering information-theoretically optimal channel codes. Calculations show that such a joint approach yields substantial improvement in QoS performance in a variety of communication scenarios.

Keywords: Information theory, random coding bound, large deviations, queuing, mobile wireless.

I. INTRODUCTION

The design of future packet cellular networks is likely to involve explicit provisioning for Quality of Service (QoS), such as delay and data rate guarantees. However, this requirement poses a challenge in wireless network design, because wireless channels have low reliability, and time varying signal strength, which may cause severe QoS violations.

In networking, QoS guarantees for delay-sensitive applications have typically been provided by the analysis of corresponding queuing systems. In [1], we considered a pure queuing model, where the effect of channel variations on link performance was captured by a single function called ‘effective capacity’. Since the focus of that paper was on the queuing model, the paper assumed ideal channel codes, so that the instantaneous capacity was assumed to be achieved at any time instant. While the paper successfully showed that the effective capacity function provides an efficient and compact representation of link performance, the results there will typically be too optimistic for a real system, which must deal with channel noise, using channel codes operating below capacity. Thus, in reality, one needs to consider not only the queuing model (as is common in networking), but also an explicit physical layer model for channel coding (requiring information theory). Then, the QoS will depend not only on delay bound violations (caused by overload of the queued server), but equally importantly, on decoding errors, introduced by the channel noise. An optimal system must, therefore, balance these two causes of errors.

To this end, consider the “joint queuing and coding system” shown in Figure 1. This joint system is fed by a source of constant rate μ . Since the channel is time-varying (with time-varying channel state), the queue attempts to match the source rate with the instantaneous quality of the channel. In Figure 1, we explicitly model a channel encoder, which potentially operates at a rate less than capacity. The encoder receives data from the queue at a time-varying rate, and must encode the data at a rate commensurate with the instantaneous channel quality. Now, if the queue has a high instantaneous output rate, then the encoder must choose channel codes with large rates, thus clearing the queue quickly, but resulting in a high decoding error probability. On the other hand, if the system chooses channel codes with low rates, it will be able to reduce the decoding error probability, only at the expense of large (potentially unstable) queuing delays. This argument shows that there must exist a system that balances the queuing and the channel coding operations optimally, resulting in the best possible QoS. This paper is an attempt to combine ideas from the fields of queuing theory and communication/information theory to design such an optimal system. For reasons of tractability, this early attempt imposes a ‘memoryless’ limitation on the server, and thus, may not be optimal. (However, we show that even this memoryless joint system achieves better performance than a pure coding system!) Under this memoryless-server condition, the paper shows that the optimal system achieves a QoS exponent (i.e., the exponent of decay of QoS violation probability with delay bound) that is the minimum of a) the delay bound violation exponent of a certain queuing system, and b) the random coding error exponent of a certain channel coding system. Since these two exponents represent the large-delay asymptotic limit of “QoS” in queuing theory and information theory respectively, the result of this paper has a satisfying interpretation.

Note that other researchers have considered either the pure queuing problem [2] or the pure channel coding problem (see references in [3]), for wireless time-varying channels. For example, a typical channel coding framework is to assume either a) the source has infinite bits in its buffer, so that the encoder can request as much data as it deems appropriate (depending only on the channel quality) at each time instant, or b) the source has a constant rate, but so does the encoder, which therefore, eliminates the need for queuing. Note that while we also assume a constant rate source (as in (b)), we explicitly allow time-varying coding rates, thus allowing for the queue in Figure 1. We then analyze the joint system in the analytically

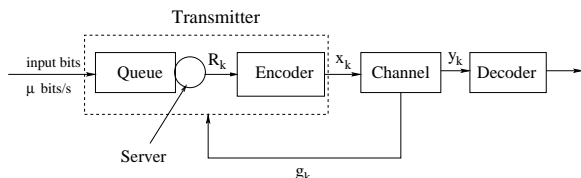


Fig. 1. Joint queuing and coding system

tractable large-delay bound regime. We must mention that there have been several excellent papers on achieving such a combination using practical channel codes [4], such as specific rate-adaptive codes, but such attempts have so far not analyzed the ultimate limits of such a joint communication system, as our paper does. Interesting combinations of information theory and queuing for multiple access was presented in [5], while [6] considered the information capacity of queues. Several papers have explored trade-offs between average power/throughput and average delay [7], [8], [9], [10]. Problems with constraint on absolute delay are considered in [7], [11]. However, each of these papers assumes that capacity achieving channel codes can be used.

Section II introduces the problem of providing QoS guarantees formally. Section III provides an overview of pure queuing theoretic and pure information theoretic asymptotic approaches in the large delay regime. It then provides the main result of this paper on the optimal joint approach. Some illustrative examples, showing the value of the joint approach are presented in Section IV. Finally, Section V concludes the paper. For readability, the main proofs are placed in the Appendix.

II. PROBLEM FORMULATION

We begin by formally describing the system model, shown in Figure 1. The discrete-time memoryless channel (DMC) model [12] is specified by the conditional probability distribution $p(y_k|g_k, x_k)$, $k = 1, 2, \dots$, where x_k, y_k are the channel input symbol and output sample respectively, at time k , while g_k is the Channel State Information CSI (e.g., gain) at time k . In simulations, we will assume that the noise is Gaussian and that Gaussian codebooks are used, but our results apply to the general DMC. The CSI is assumed to be known perfectly to the transmitter (so that it can adapt to the CSI) as well as to the receiver (so that it can use maximum-likelihood (ML) decoding.) For simplicity of presentation, we first analyze the case with independent, identically distributed (i.i.d.) g_k and then trivially extend our result to block fading channels [13], which have blocks of equal gain.

Notice that the transmitter consists of a queue+server followed by an encoder. Thus, it combines the canonical models of queuing and information theory. Throughout the paper, we measure information in ‘nats’, instead of bits (i.e., $\log_e(\cdot)$ assumed for entropy) for convenience. We assume that the server chooses an ‘instantaneous server capacity’ R_k nats, which depends only on the current CSI. i.e., $R_k = R(g_k)$ for some fixed function $R(g)$. Thus, the server rate is assumed to be a memoryless function of channel CSI. (In the rest of the paper, an ‘optimal joint system’ refers to the optimal system, under the memoryless-server assumption.) The encoder receives data from the server at rate R_k , encodes the data

bits into a sequence of symbols and transmits the channel code sequentially into the channel. We assume a streaming-code encoder (such as a convolutional encoder) [14], since a block-encoder [12], although simpler, requires additional delay to buffer the data until a block of data is ready for encoding. Thus, we assume that bits that depart the server at time $(k-1)$ are encoded into symbols x_k, x_{k+1}, \dots

The decoder decodes the channel code after buffering the received y_k sufficiently. The system needs to be designed so that the QoS demanded by the source application is met. We formally define QoS as follows. The source has a constant rate μ and a maximum delay bound of D . Given the channel statistics and the system design, it is required that the probability of error be as small as possible. (Alternatively, the problem can also be cast in the framework of specifying the channel resource that results in a certain maximum tolerable probability of error.) The probability of error must capture both; delay bound violations due to queuing, as well as decoding errors due to channel noise. This will be specified in the following.

Consider the source bit that arrives at the queue at time $k - D_k$. After spending time D_k in the queue, it departs the server at time $k - 1$. The encoder encodes the bit(s) into the code symbols x_k, x_{k+1}, \dots . Since the source demands a delay bound of D , the decoder must decode this bit at time $k - D_k + D - 1$. Let P_{err} be the probability of bit decoding error. Then, a bit decoding error can occur in two cases; a) the bit spent the entire time D (or more) in the queue, and therefore, never got sent, or b) the bit was sent within a delay of D , but was decoded incorrectly (either because it spent too much time in the queue, leaving too little time for the channel code to be effective, or because of unusually large channel noise). Thus, P_{err} captures both, the queuing notion of error (delay-bound violation) and the information-theoretic notion of decoding error. Therefore, the problem is to design a joint system that can minimize P_{err} for the given μ, D and channel statistics. Note that our formulation encompasses a pure coding framework, which does not allow queuing, and thus, is expected to perform better than the latter.

The main result of this paper appears next.

III. JOINT QUEUEING AND CODING

A. Background

The key insight of this paper is as follows. If the encoder is assumed to use ‘ideal channel’ codes that achieve the instantaneous capacity, then the recently developed effective capacity [1] result can be used to calculate the P_{err} (which is simply the delay bound violation probability now) in the asymptotic regime of $D \rightarrow \infty$. Essentially, effective capacity is a pure queuing (i.e., no coding) large-deviations framework, which is the channel dual of the notion of effective bandwidth, developed to quantify source burstiness in Asynchronous Transfer Mode (ATM) networks. To elaborate, for a channel whose instantaneous capacity at time t is $r(t)$ (recall that capacity-achieving codes are assumed in a pure queuing approach), it was shown [1] that if $P_{err} \doteq \Pr\{D(\infty) \geq D\}$ (i.e., delay

bound D violation probability at steady state), then

$$\lim_{D \rightarrow \infty} \frac{-1}{D} \log(P_{err}) = \theta_q(\mu) \quad (1)$$

where $\theta_q(\mu) = \mu \alpha^{-1}(\mu)$ and $\alpha^{-1}(\cdot)$ is the inverse mapping of the following (what we call) effective capacity function (if it exists),

$$\alpha(u) = - \lim_{t \rightarrow \infty} \frac{1}{ut} \log \mathbf{E}[e^{-u \sum_{\tau=0}^{t-1} r(\tau)}], \quad \forall u \geq 0. \quad (2)$$

The effective capacity function exists for a wide range of channel processes; in particular for i.i.d. CSI, it simply reduces to the ratio of log-moment generating function of the instantaneous capacity to the exponent u (see eqn. (34)). In this paper, we will not require the general form (2) of the effective capacity function, since we restrict ourselves to the i.i.d. (block-fading) case only. However, interestingly, we will show shortly that the effective capacity function appears as one of the components of QoS in the joint queuing/coding case.

On the other hand, if a pure coding approach is considered (i.e., no queuing, or server rate $R_k \equiv \mu, \forall k$), then the random coding exponent [12] provides a bound on P_{err} (which is simply the decoding error probability now). To elaborate, for a block code of rate μ and code length $D/2$, the decoding error probability can be bounded as,

$$P_{err} \leq \exp(-(D/2)E_c(\mu)) \quad (3)$$

$$E_c(\mu) = \max_{0 \leq \rho \leq 1} E_0(\rho, \mu) \quad \text{'error exponent'}. \quad (4)$$

For example, for Gaussian noise and codebooks,

$$E_0(\rho, \mu) = - \log \mathbf{E} \exp \left\{ -\rho \left(\log \left(1 + \frac{gP_0}{1+\rho} \right) - \mu \right) \right\} \quad (5)$$

where P_0 is the ratio of transmit power to noise power. Notice how $\theta_q(\mu)$ and $E_c(\mu)$ are both exponents that bound P_{err} , using two completely different approaches, which apply respectively, under two different assumptions. Essentially, our main contribution in this paper is to show that the optimal joint queuing/coding system considers both these exponents, to obtain the best possible error exponent for P_{err} .

B. A joint queuing/coding exponent

Since we have limited our design to joint queuing/coding systems which have a server 'instantaneous capacity' $R_k = R(g_k)$ (i.e., memoryless), we need to choose the function $R(g) \geq 0$ correctly. This choice must be made so that P_{err} is minimized for the given μ, D . We work in the regime of asymptotically large D , so that large-deviations queuing and information theoretic coding results can be applied. Thus, the problem is to choose $R(g)$ so that the 'joint exponent' $\theta(\mu)$ is maximized, where

$$\liminf_{D \rightarrow \infty} \frac{-1}{D} \log(P_{err}) = \theta(\mu). \quad (6)$$

In order to calculate the joint exponent, we need to analyze the combined queuing and coding system. This analysis is done in two stages. In stage 1, for a *fixed decoding delay*, analyze the decoding error probability of the streaming code, which encodes bits received by the encoder at time k , into

symbols x_k, x_{k+1}, \dots , possibly encoding a different number of new bits in different symbols (as specified by the server capacity $R_k = R(g_k)$). Note that we use a streaming code, rather than a block code, to avoid the extra delay in the latter, whose encoding delay equals decoding delay. If the specified decoding delay is zero or less, we can upper bound the error probability by one. In stage 2, we analyze the queuing delay D_k caused by the queuing system (due to the chosen $R(g)$), and thus, obtain the decoding delay $D - D_k$ available for the decoder, for the QoS delay bound to be met. For bits that have $D - D_k \leq 0$, we bound the error probability by one, since these bits never get transmitted. Thus, the *bit error probability* P_{err} of the code, fed by the queue, can be calculated by combining stages 1 and 2.

The result of the analysis is formalized in the following two lemmas.

Lemma 1: For a streaming code with a delay-bound of D , and source rate $R_k = R(g_k)$, the error exponent is given by,

$$\liminf_{D \rightarrow \infty} \frac{-1}{D} \log(P_{err}) \geq E_c \quad (7)$$

where E_c is given by (20). Note that this is the same exponent as achieved by a block code, whose block length is equal to D .

Proof: The proof follows [12], adapted to the case of a streaming code, and assuming the specific time-varying source rate. See Appendix A for details. This is the analysis for stage 1. ■

Lemma 2: Consider a joint queuing/coding system (Figure 1), which has a constant source rate μ , a delay-bound of D , and which employs a streaming code. If the channel is a fading channel with i.i.d. symbol-by-symbol fading, the error exponent is given by,

$$\liminf_{D \rightarrow \infty} \frac{-1}{D} \log(P_{err}) \geq \theta^* \quad (8)$$

where θ^* is given by (35).

Proof: The proof proceeds by analyzing the queuing delay (stage 2), and then uses Lemma 1 to analyze the joint system. See Appendix B for details. ■

Lemma 2 shows that the optimum error exponent for the joint queuing/coding system is given by the minimum of the respective error exponents $\theta_q(\mu)$ and $E_c(\mu)$ (see (35)). Thus, the joint system must balance the requirements of queuing (choosing large $R(g)$ so that the queue clears quickly) with the requirements of the encoder (choosing small $R(g)$ so that the code is highly redundant, and therefore, not susceptible to errors due to noise). Thus, the lemma provides a satisfyingly symmetric result that clearly shows the trade-off between queuing and coding.

The following lemma extends the scope of Lemma 2, by considering a block-fading channel model, where the CSI is constant over a block of length T symbols, but i.i.d. across blocks.

Lemma 3: Consider a joint queuing/coding system (Figure 1), which has a delay bound of D , and which employs a streaming code. If the channel is i.i.d. block-fading with block length T , and constant channel gain over each block, the error

exponent is given by,

$$\liminf_{D \rightarrow \infty} \frac{-1}{D} \log(P_{err}) \geq \theta^* \quad (9)$$

where θ^* is given by (40).

Proof: The correlation of CSI within each block introduces an intractable dependency between the queue and the encoder. The proof adapts Lemma 2, by neglecting at most $2T$ ($\ll D$) symbols which cause the dependency. See Appendix C for details. ■

In the next section, we show that the joint queuing and coding approach provides a significant QoS improvement, as evidenced by a large error exponent θ^* , in a range of situations.

IV. ILLUSTRATIVE RESULTS

We compare the joint coding/queuing system with a pure queuing system (which assumes that the instantaneous capacity can be achieved) and the pure coding system (which sets $R(g) \equiv \mu$, indicating the absence of a queue). Since the queuing system does not use channel coding, if noise is present in the channel, it will be unable to achieve arbitrarily small P_{err} , even at large D . Therefore, we assume that the only source of error for the queuing system is delay-bound violation. Consequently, the error exponent $\theta_q(\mu)$ that we obtain for the queuing system should only be interpreted as an upper bound on the maximum possible joint exponent $\theta^*(\mu)$. $\theta_q(\mu)$ will not be achievable in a real system.

We will only present toy results in this paper to illustrate the potential utility of the joint approach. Thus, assume a simple Gaussian channel model, $y_k = \sqrt{g_k}x_k + n_k$, where the noise n_k is Gaussian with variance one, and Gaussian codebooks, transmitted at fixed power P_0 per-symbol, are used. The channel gain is assumed to be binary. i.e., $g_k \in \{0, g_{max}\}$. A range of (constant) source rate μ , average signal-to-noise-ratio $SNR = P_0 \mathbf{E}[g]$ and fading block-lengths T are simulated. $\theta_q(\mu)$, $E_c(\mu)$, $\theta^*(\mu)$ are calculated and shown below. Note that $P_{err} \approx \exp(-\theta \cdot D)$ holds for each exponent, respectively.

The optimization problem (35) can be solved by a brute-force search over the $R(g)$ function as follows. For the given μ , T and channel probability mass function $p(g)$, fix $R(g) \geq 0$. Calculate (20) by maximizing over ρ . This can be done efficiently, since it is a convex problem. Now, set $\theta = E_c$ and mark θ as feasible if $\alpha_R(E_c/\mu) > \mu$. This evaluation can be done by calculating the log-moment generating function (34). Repeat the procedure over a suitably discretized search space of $R(g)$ and obtain the maximum θ . For the pure coding exponent $E_c(\mu)$, simply set $R(g) = \mu \forall g$. The pure queuing exponent $\theta_q(\mu)$ is obtained as $\mu \alpha^{-1}(\mu)$. For more general channel statistics $p(g)$, the brute-force search may be too expensive. However, the optimization can still be done efficiently because it can be cast as a quasi-convex problem [15]. Details will appear in a more extensive journal publication, currently under preparation.

Figures 2, 3 and 4 show that the pure queuing exponent is an upper bound on both, the joint queuing/coding and pure coding exponents. These figures show that the joint queuing/coding system provides a substantial gain over the pure coding system for a wide range of μ , average SNR and T .

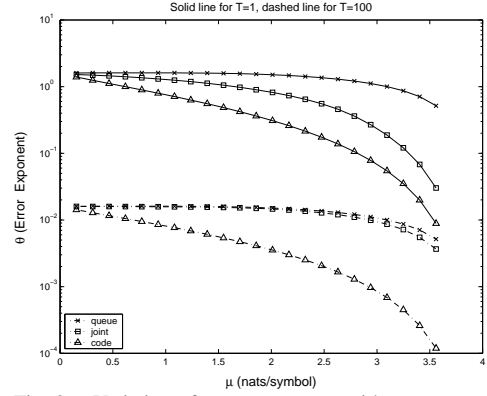


Fig. 2. Variation of error exponents with source rate

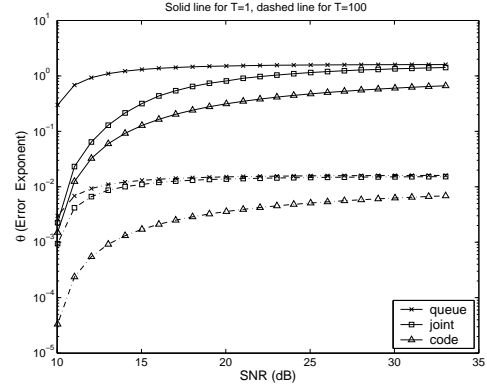


Fig. 3. Variation of error exponents with SNR

Figure 2 shows the variation of error exponents with source rate at $SNR = 20$ dB and $T = 1, 100$. As μ increases, the QoS requirement becomes tighter, thus reducing the error exponents. The average Shannon capacity in this case is 3.87 nats/symbol. Observe that the error exponents go to zero for source rates greater than the average capacity.

Figure 3 shows the variation of error exponents with average SNR at $\mu = 2$ nats/symbol and $T = 1, 100$. The plot shows that the pure coding and joint queuing/coding systems perform much worse than the pure queuing system at low SNR . However, a practical queuing system will be hard hit by noise at low SNR . This fact is not reflected in the θ_q plot, since we (naively) assume that the pure queuing system achieves the instantaneous capacity. This figure brings out the fact that the joint system gives a substantial gain over the pure coding system only at high values of SNR for $T = 1$. This is because under very noisy conditions, most of the delay is allotted to the code, leaving very little for the queue. However, for higher values of T (as is likely for typical fading scenarios), the performance of the joint system is close to the pure queuing system.

Figure 4 shows the variation of error exponents with block length T at $\mu = 2$ nats/symbol and $SNR = 10, 15$ dB. We observe that the performance of the joint queuing/coding system is very close to the pure queuing system for large values of T . The performance difference between the joint queuing/coding and pure coding systems increases with T .

V. CONCLUSION

This paper presented a joint approach to queuing and channel coding, by combining ideas from large deviation

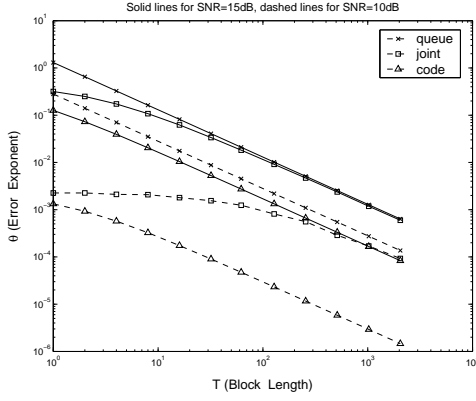


Fig. 4. Variation of error exponents with block length

theory (effective capacity) and information theory (random coding bounds), which are applicable in the regime of large delay-bounds ($D \rightarrow \infty$). This approach aims to find the ultimate (asymptotic) limit of delay-bounded communication, in contrast to other approaches, which attempt such a combination using specific queuing and practical channel coding models, such as rate-adaptive convolutional coding. As a first step, the paper considered a memoryless server model, where the instantaneous server capacity was chosen to be a function $R(g)$ of current CSI. A lower bound on the decay exponent of decoding error probability was calculated for a joint queuing/coding system, assuming i.i.d. block fading. This exponent cannot be worse than the exponent in a pure coding system (while the pure queuing system fails in the absence of channel coding), since setting $R(g) \equiv \mu$ reduces the joint system to the pure coding system. Simulation results show that the joint approach can provide significantly better QoS, as shown by a larger error exponent, over a range of situations, as compared to the pure coding approach. Note, however, that a pure coding approach does not require CSI at the transmitter, unlike the joint approach. Future work will attempt to extend the approach to more general channel and joint system models, and design practical streaming codes which utilize the asymptotic theory.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under the grant ANI-0111818.

APPENDIX

A. Proof of Lemma 1

We denote the sequence $h_1^k = \{h_1, \dots, h_k\}$. $\mathbf{E}_{\mathbf{g}}[\cdot]$ denotes expectation over \mathbf{g} .

Encoding: We denote the symbol transmitted at time k by x_k and the set of bits that arrive at the encoder over time $1, 2, \dots, k$ as $\bar{b}(k)$. We assume that $R_k = |\bar{b}(k)| - |\bar{b}(k-1)|$ is a time-varying source rate, which depends solely on the current CSI g_k as $R_k = R(g_k)$. We consider a streaming code where x_k is generated based on $\bar{b}(k)$, i.e., $x_k = f(\bar{b}(k))$ where, $f(\cdot)$ is a probabilistic function. Specifically, for each $\bar{b}(k)$, a symbol x_k is chosen according to an arbitrary probability distribution $Q(x)$, independent of $x_{k-1}, x_{k-2}, \dots, x_1$. Notice that two bit streams will have the same code symbols prior

to the time at which *they first differ* from each other, and will have independent symbols from that time onwards. The probability of choosing a particular codeword x_1^{k+D-1} in the random code generation is

$$Q(x_1^{k+D-1}) = \prod_{i=1}^{k+D-1} Q(x_i). \quad (10)$$

Decoding: We consider maximum-likelihood decoding, assuming that the decoder knows the CSI. The k^{th} symbol is decoded at time $k + D - 1$ (since a delay constraint of D symbols is assumed). To decode the symbol x_k , all samples y_1^{k+D-1} are utilized. We now compute the probability that the k^{th} symbol is incorrectly decoded, given that message $m = \bar{b}(k + D - 1)$ is transmitted. Notice that the message is assumed to consist of bits up to time $k + D - 1$, since those time samples are used for decoding x_k . Most of the proof follows [12], so we skip several steps and retain the ones that highlight the difference between a streaming and block code. The probability of decoding the k^{th} symbol incorrectly, averaged over the random code, is given as an expectation of the probability over the CSI sequence \mathbf{g} as,

$$\begin{aligned} P_{e,m} &= \mathbf{E}_{\mathbf{g}}[P_{e,m}(\mathbf{g})] = \mathbf{E}_{\mathbf{g}}[P(\hat{x}_k \neq x_k | m)] \\ &= \mathbf{E}_{\mathbf{g}} \left[\sum_{x_1^{k+D-1}} \sum_{y_1^{k+D-1}} P(y_1^{k+D-1} | x_1^{k+D-1}) Q(x_1^{k+D-1}) \right. \\ &\quad \left. P(\hat{x}_k \neq x_k | y_1^{k+D-1}, x_1^{k+D-1}, m) \right]. \end{aligned} \quad (11)$$

$$\begin{aligned} \text{But, } P(\hat{x}_k \neq x_k | y_1^{k+D-1}, x_1^{k+D-1}, m) \\ \leq \left(\sum_{\{\tilde{m}: \tilde{x}_k \neq x_k\}} P(A_{\tilde{m}}) \right)^\rho, \quad 0 \leq \rho \leq 1, \end{aligned} \quad (12)$$

where $A_{\tilde{m}}$ is the event that the decoded message is \tilde{m} , conditioned on y_1^{k+D-1}, x_1^{k+D-1} and m . Note that only such \tilde{m} are considered which result in $\tilde{x}_k \neq x_k$. Let $\tilde{b}(k + D - 1)$ denote the bits of \tilde{m} until time $k + D - 1$.

$$\begin{aligned} P(A_{\tilde{m}}) &\leq \sum_{\tilde{x}_1^{k+D-1}} Q(\tilde{x}_1^{k+D-1} | \tilde{m}, x_1^{k+D-1}, m) \\ &\quad \left(\frac{P(y_1^{k+D-1} | \tilde{x}_1^{k+D-1})}{P(y_1^{k+D-1} | x_1^{k+D-1})} \right)^s, \quad s \geq 0 \end{aligned} \quad (13)$$

From (12) and (13) we obtain,

$$\begin{aligned} P(\hat{x}_k \neq x_k | y_1^{k+D-1}, x_1^{k+D-1}, m) &= \\ \left[\sum_{\{\tilde{m}: \tilde{x}_k \neq x_k\}} \left\{ \sum_{\tilde{x}_1^{k+D-1}} Q(\tilde{x}_1^{k+D-1} | \tilde{m}, x_1^{k+D-1}, m) \right. \right. \\ &\quad \left. \left. \left(\frac{P(y_1^{k+D-1} | \tilde{x}_1^{k+D-1})}{P(y_1^{k+D-1} | x_1^{k+D-1})} \right)^s \right\}^\rho \right]. \end{aligned} \quad (14)$$

We partition the codewords $\{\tilde{m} : \tilde{x}_k \neq x_k\}$ into $k + 1$ subsets as follows. Let $\{\tilde{x}_j\}$ be the symbols of \tilde{m} .

$$\begin{aligned} B_p &= \{\tilde{m} : \tilde{b}(j) = \bar{b}(j) \quad \forall j < p, \quad \tilde{b}(p) \neq \bar{b}(p)\} \\ |B_p| &\leq \exp \left(\sum_{i=p}^{k+D-1} R_i \right) \end{aligned} \quad (15)$$

Thus, $\tilde{m} \in B_p$ implies that m and \tilde{m} have the same bits in the first $p - 1$ time instants, while at least one of the bits at

time p (and perhaps more in the future) are different. Since $\tilde{m} \in B_p, \Rightarrow \tilde{x}_j = x_j, j < p$, while $\tilde{x}_j \perp x_j, j \geq p$ (since the code generates independent symbols for different messages), hence,

$$Q(\tilde{x}_1^{k+D-1} | \tilde{m}, x_1^{k+D-1}, m) = \prod_{i=1}^{p-1} \delta(\tilde{x}_i - x_i) \prod_{i=p}^{k+D-1} Q(\tilde{x}_i). \quad (16)$$

We only need to consider the messages \tilde{m} for which at most the first $(k-1)$ symbols are the same (i.e., $p \leq k$). If more than the first $(k-1)$ symbols are same, then there is no error in decoding x_k . Now, using (15) and (16), (14) simplifies to

$$\begin{aligned} P(\hat{x}_k \neq x_k | y_1^{k+D-1}, x_1^{k+D-1}, m) &\leq \\ &\left[\sum_{p=0}^k \exp\left(\sum_{i=p}^{k+D-1} R_i\right) \sum_{\tilde{x}_1^{k+D-1}} \left\{ \prod_{i=1}^{p-1} \delta(\tilde{x}_i - x_i) \right\} \left\{ \prod_{i=p}^{k+D-1} Q(\tilde{x}_i) \right\} \right. \\ &\quad \left. \left(\frac{P(y_1^{k+D-1} | \tilde{x}_1^{k+D-1})}{P(y_1^{k+D-1} | x_1^{k+D-1})} \right)^{s\rho} \right] \\ &\stackrel{a}{=} \left[\sum_{p=0}^k \exp\left(\sum_{i=p}^{k+D-1} R_i\right) \left\{ \prod_{i=p}^{k+D-1} \left(\sum_{\tilde{x}_i} Q(\tilde{x}_i) \left(\frac{P(y_i | \tilde{x}_i)}{P(y_i | x_i)} \right)^s \right) \right\} \right]^\rho \end{aligned}$$

where, (a) is because the channel is a DMC.

Applying $(\sum_i \lambda_i)^\rho \leq \sum_i \lambda_i^\rho$, if $\lambda_i \geq 0, \rho \leq 1$ in (11),

$$\begin{aligned} P_{e,m} &\leq \mathbf{E}_{\mathbf{g}} \left[\sum_{y_1^{k+D-1}} \sum_{x_1^{k+D-1}} P(y_1^{k+D-1} | x_1^{k+D-1}) Q(x_1^{k+D-1}) \right. \\ &\quad \left. \left[\sum_{p=0}^k \exp\left(\sum_{i=p}^{k+D-1} \rho R_i\right) \left\{ \prod_{i=p}^{k+D-1} \left(\sum_{\tilde{x}_i} Q(\tilde{x}_i) \left(\frac{P(y_i | \tilde{x}_i)}{P(y_i | x_i)} \right)^s \right) \right\} \right]^\rho \right] \\ &= \mathbf{E}_{\mathbf{g}} \left[\sum_{p=0}^k \exp\left(\sum_{i=p}^{k+D-1} \rho R_i\right) \sum_{y_1^{k+D-1}} \left[\left\{ \prod_{i=1}^{p-1} \left(\sum_{x_i} Q(x_i) \right. \right. \right. \right. \\ &\quad \left. \left. \left. P(y_i | x_i) \right\} \left\{ \prod_{i=p}^{k+D-1} \left(\sum_{x_i} Q(x_i) P(y_i | x_i)^{1-s\rho} \right) \right\} \right. \right. \\ &\quad \left. \left. \left[\prod_{i=p}^{k+D-1} \left(\sum_{x_i} Q(x_i) P(y_i | x_i)^s \right) \right]^\rho \right] \right]. \end{aligned}$$

It can be shown that $s = 1/(1+\rho)$ minimizes the above expression [12]. Therefore, we substitute $1/(1+\rho)$ for s to obtain the tightest bound.

$$\begin{aligned} P_{e,m} &\leq \mathbf{E}_{\mathbf{g}} \left[\sum_{p=0}^k \exp\left(\sum_{i=p}^{k+D-1} \rho R_i\right) \right. \\ &\quad \left. \prod_{i=p}^{k+D-1} \left(\sum_{y_i} \left(\sum_{x_i} Q(x_i) P(y_i | x_i)^{1/(1+\rho)} \right)^{1+\rho} \right) \right] \\ &= \mathbf{E}_{\mathbf{g}} \left[\sum_{p=0}^k \exp\left\{-\sum_{i=p}^{k+D-1} (E(g_i) - \rho R(g_i))\right\} \right] \quad (17) \end{aligned}$$

where $E(g_i)$ depends on the current channel transition matrix $P(y|x, g_i)$ as below,

$$E(g_i) = -\log \left[\sum_y \left(\sum_x Q(x) P(y|x, g_i)^{1/(1+\rho)} \right)^{1+\rho} \right]$$

Since $E(g_i), R(g_i)$ depend only on the current g_i , and the g_i 's are i.i.d., choosing the tightest bound gives,

$$P_{e,m} \leq \sum_{p=0}^k \exp\{-(k+D-p) \cdot E_c\} \quad (18)$$

$$\leq c_1 \exp(-D \cdot E_c) \quad \text{where} \quad (19)$$

$$E_c = \max_{0 \leq \rho \leq 1} (-\log(\mathbf{E}_{\mathbf{g}}[\exp(R(g) - E(g))])) \quad (20)$$

assuming $E_c > 0$ by appropriate choice of $R(g)$, and c_1 is some constant. Since any bit at time k is decoded incorrectly only if x_k is decoded incorrectly, the bit error probability $P_{err} \leq P_{e,m}$ above.

B. Proof of Lemma 2

Let g_k be the i.i.d. channel state. Let q_k be the queue length at time k and D_k be the queuing delay of the bit that departs the queue at time k . Since the source rate is assumed constant, $D_k = q_k/\mu$ [1]. We choose the server rate to be a memoryless function $R_k = R(g_k)$, where $R(g)$ is a fixed function. q_{k+1} can be expanded as [16],

$$q_{k+1} = \max(0, \mu - R_{k+1}, 2\mu - R_{k+1} - R_k, 3\mu - R_{k+1} - R_k - R_{k-1}, \dots) \quad (21)$$

$$D_{k+1} = \max(0, a_{k+1}, a_{k+1} + a_k, \dots) \quad (22)$$

$$\text{where, } a_k \doteq 1 - R_k/\mu = 1 - R(g_k)/\mu. \quad (23)$$

The bit error probability P_{err} can be bounded as,

$$P_{err} \leq \Pr(\hat{x}_k \neq x_k) = \mathbf{E}_{\mathbf{g}}[P_{e,m}(\mathbf{g})] \quad (24)$$

$$= \mathbf{E}_{D_k}[\mathbf{E}_{\mathbf{g}}[P_{e,m}(\mathbf{g}) | D_k]] \quad (25)$$

$$\stackrel{a}{\leq} \mathbf{E}_{D_k}[c_1 \cdot \exp(-(D - D_k)E_c)] \quad (26)$$

$$= c_1 \cdot \exp(-DE_c) \mathbf{E}_{D_k}[\exp(D_k \cdot E_c)] \quad (27)$$

where (a) arises because of the following. The time D_k spent by the bit in the queue is a function of g_k, g_{k-1}, \dots while the decoding error probability, given D_k , depends on g_{k+1}, g_{k+2}, \dots . Since g_i 's are i.i.d., the distribution of g_{k+1}, g_{k+2}, \dots , given D_k , remains the same. Further, $R_i = R(g_i)$, so the conditions of Lemma 1 are satisfied, with the decoder allowed a (remaining) delay constraint of $D - D_k$. $R(g)$ must be chosen, so that E_c (defined by (20)) satisfies $E_c > 0$.

$$\begin{aligned} \exp(D_k E_c) &= \exp(E_c \cdot \max(0, a_k, a_k + a_{k-1}, \dots)) \\ &\leq \sum_{j=-1}^{\infty} \exp(E_c \sum_{i=0}^j a_{k-i}) \end{aligned}$$

where the summand is defined as 1 for $j = -1$.

$$P_{err} \leq c_1 \cdot \exp(-DE_c) \left[\sum_{j=0}^{\infty} (\mathbf{E} \exp(E_c a))^j \right] \quad (28)$$

since a_k are i.i.d. (23). Choose E_c so that

$$\mathbf{E} \exp(E_c a) < 1 \quad (29)$$

$$\text{Then, } P_{err} \leq c_3 \cdot \exp(-D \cdot E_c) \quad (30)$$

$$\begin{aligned} \text{Define joint exponent } \theta &\doteq \liminf_{D \rightarrow \infty} -\frac{1}{D} \log P_{err} \\ &\geq E_c \quad (\text{using (30)}) \end{aligned}$$

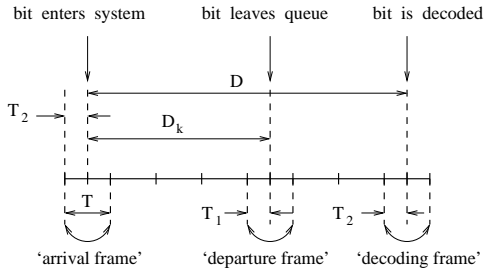


Fig. 5. Block fading channel

Therefore, a lower bound on the optimum joint exponent is obtained by solving the following optimization problem,

$$\theta^*(\mu) = \max_{R(g) \geq 0} \theta \quad (31)$$

$$\text{s.t. } \theta < E_c, \quad \mathbf{E} \exp(E_c a) < 1. \quad (32)$$

Eqn. (32) can be written as

$$\alpha_R(E_c/\mu) > \mu \quad (33)$$

$$\text{where, } \alpha_R(\phi) = -\frac{1}{\phi} \log[\mathbf{E} \exp(-\phi R(g))] \quad (34)$$

is the effective capacity function corresponding to $R(g)$. Therefore, we require $E_c(\mu) \doteq E_c < \mu \alpha_R^{-1}(\mu) \doteq \theta_q(\mu)$ since $\alpha_R(\phi)$ is a decreasing function [17]. Since scaling $R(g)$ up strictly increases θ_q (to an arbitrarily large value) while strictly decreasing $E_c(\mu)$ (up to the lower bound of zero), the optimization problem (31) can be stated compactly as,

$$\theta^*(\mu) = \max_{R(g) \geq 0} \min\{E_c(\mu), \theta_q(\mu)\} \quad (35)$$

where, $E_c(\mu)$ is given by (20), while $\theta_q(\mu) = \mu \alpha_R^{-1}(\mu)$, obtained using (34).

C. Proof of Lemma 3

We consider a block fading channel with i.i.d. gains and block length T of each block. Let the gain for the k^{th} symbol x_k be g_k . Then, $g_i = g_j$ if $\lfloor i/T \rfloor = \lfloor j/T \rfloor$. We assume a delay constraint $D = lT$, where l is a positive integer. Let the arrival-instant of the bit of interest to the queue, be offset from the start boundary of its arrival frame by T_2 and the instant of departure from the queue be offset from the start boundary of the departure frame by T_1 (see Figure 5).

In this case, the evaluation steps for P_{err} are similar to those for the $T = 1$ case analyzed in Appendix B. We note the steps that differ. The only complication here is that the bit may be sent within a frame rather than at its boundary, thus voiding the assumption of independence of $\{g_{k+1}, g_{k+2}, \dots\}$ and $\{g_k, g_{k-1}, \dots\}$. But this can be handled easily by throwing away all the symbols within the departure frame, as well as all the symbols within the last frame used for decoding, from any decoding considerations. There are at most $2T \ll D$ such symbols. Therefore, this can only increase the error probability by a constant factor (i.e., constant with respect to D), which does not affect the decoding error exponent, in the limit of large D . Thus, following Appendix B, we have similar to (27),

$$P_{err} \leq c_2 \exp\{-DE_c\} \cdot \mathbf{E}_g[\exp\{D_k E_c\}], \quad (36)$$

$$E_c^T \doteq \max_{0 \leq \rho \leq 1} \left(-\frac{1}{T} \log(\mathbf{E}_g[\exp(T \cdot (R(g) - E(g)))] \right) \quad (37)$$

where, c_2 is a constant, which accounts for the $2T$ symbols thrown away. Note the difference between (20) and (37), due to the correlation of g_i within each frame. Similar to (28), but since g_k is constant within a frame, the expectation term in (36) can be bounded as,

$$\mathbf{E}_{D_k}[\exp\{D_k E_c\}] \leq (T_1 + 1) \max(1, \mathbf{E}[\exp\{E_c a_k T_1\}]) + \sum_{j=T_1+1}^{\infty} \exp\{E_c(a_k + a_{k-1} + \dots + a_{k-j})\}. \quad (38)$$

The second term in the above equation is similar to the term in (28). Similar to (29), we assume that $\mathbf{E} \exp(E_c a T) < 1$ holds. Thus, the error probability is bounded by,

$$P_{err} \leq c_4 \exp\{-DE_c\} \quad \text{where, } c_4 \text{ is a constant.} \quad (39)$$

Thus, the optimum joint error exponent $\theta^{*T}(\mu)$ can be found as below,

$$\theta^{*T}(\mu) = \max_{R(g) \geq 0} \min\{E_c^T(\mu), \theta_q^T(\mu)\} \quad (40)$$

where, $E_c^T(\mu)$ is given by (37), while $\theta_q^T(\mu) = (\mu/T) \alpha_R^{-1}(\mu)$, obtained using (34).

REFERENCES

- [1] Dapeng Wu, R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Communications*, pp. 630-643, July 2003.
- [2] G. Kesidis, J. Walrand, and C. Chang, "Effective bandwidths for multiclass markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424-428, Aug. 1993.
- [3] R. Negi, "Power adaptation strategies for delay constrained channels," *Ph.D. thesis*, Stanford University, 2000.
- [4] J. Hagenauer, N. Seshadri, and C. Sundberg, "The performance of rate-compatible punctured convolutional codes for digital mobile radio," *IEEE Trans. Commun.*, vol. 38, no. 7, pp. 966-980, July 1990.
- [5] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 963-969, Aug. 1995.
- [6] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 4-18, Jan. 1996.
- [7] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Info. Theory*, vol. 48, pp. 1135-1149, May 2002.
- [8] D. Rajan, A. Sabharwal and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Info. Theory*, vol. 50, pp. 125-144, January 2004.
- [9] W. S. Yoon and T. E. Klein, "Delay-optimal power control for wireless data users with average power constraint," *Proc. IEEE ISIT*, pp. 53, July 2002.
- [10] E. M. Yeh and A. S. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," *Proc. CISS*, March 2004.
- [11] R. Negi, and J. Cioffi, "Delay-constrained capacity with causal feedback," *IEEE Trans. Info. Theory*, vol. 48, pp. 2478-2494, Sept. 2002.
- [12] R. G. Gallager, "Information Theory and Reliable Communication," John Wiley and Sons, 1968.
- [13] R. J. McEliece, W. E. Stark, "Channel with block interference," *IEEE Trans. Info. Theory*, vol. 30, pp. 44-53, Jan. 1984.
- [14] E. Martinian, G. W. Wornell, "Universal Codes For Minimizing Per-User Delay on Streaming Broadcast Channels," *Proc. Allerton Conf. Commun., Contr., and Computing*, Oct. 2003.
- [15] Stephen Boyd, Lieven Vandenberghe, "Convex optimization," 1999, Course Notes, EE364, Stanford University.
- [16] L. Kleinrock, "Queueing systems: Vol. I," John Wiley, New York, 1976.
- [17] Dapeng Wu, "Providing Quality-of-Service Guarantees in Wireless Networks," *PhD thesis*, Dept. of Elec. and Comp. Eng., Carnegie Mellon University, August 2003.