# Large Deviations on Empirical Service for Erasure Channels with Memory

Santhosh Kumar, Jean-Francois Chamberland and Henry D. Pfister
Department of Electrical and Computer Engineering, Texas A&M University

***Abstract*—This article examines the performance of a digital communication link from a large deviations perspective. The underlying physical environment is modeled as an erasure channel with memory. Information is shielded from symbol erasures using error-correcting codes with finite block-lengths. At the onset of the communication process, the transmit buffer is assumed to contain a certain number of data bits; these bits are partitioned into segments and subsequently transmitted as coded packets. Acknowledgments of successful transmissions are obtained through periodic, reliable feedback. Performance criteria are derived based on the average service rate and the first-passage time to an empty queue. The optimization problem is posed in an asymptotic setting and large deviation principles are obtained for these two quantities. The proposed analysis framework provides a methodology tailored to code rate and block-length selection for delay-sensitive applications. This study leads to pertinent guidelines on how to choose system parameters for communication over correlated channels. Examples obtained through a numerical study are included to further illustrate the value of the techniques introduced in this paper.**

## I. Introduction

As mobile devices become more powerful, developers build increasingly sophisticated applications. The needs of emerging software can vary significantly in terms of rate requirements and delay tolerance. Furthermore, the optimal allocation of system resources at the physical and link layers can be highly dependent on the demands of these applications and the profiles of the underlying channels. For instance, services that are delay tolerant allow the use of interleaving, long error-correcting codes and rate adaptation. In contrast, data flows with acute delay sensitivity cannot sustain coding over extended periods of time; this precludes the use of long codewords. These contrasting facts also reflects the tension between the insights provided by classical information theory and design decisions required for implementation of contemporary cellular systems.

One factor that can greatly affect the performance of real-time communication systems is channel memory. A strong, positive correlation over time in a fading environment prolongs the average duration of deep fades and hence adversely affects performance. This phenomenon can be studied using various measures of performance, including outage capacity [1], [2] and dispersion [3]. For applications that require reliable and ordered delivery of data streams, successive transmission failures engender queue buildup. The ability of the system to recover from such fading events depends on the excess capacity of the wireless connection; draining the queue may, in many situations, take a long time. This brief discussion points to the importance of studying and understanding the queued behavior of communication systems, especially in the context of channels with memory. Elaborate treatments of the interplay between delay, queueing and communication over unreliable channels are available in the literature [4], [5], [6]. We refer the interested reader to these existing contributions and turn to the specifics of the problem we wish to address.

This article focuses on finite-state channels with memory, a popular class of models often employed to design and test communication schemes for delay-sensitive systems [7], [8], [9]. Such models enable the optimization of system parameters, such as code rate and block-length, to enhance the queued performance of delay-sensitive wireless connections. For instance, this framework can be leveraged to obtain the distribution of the first-passage time to an empty queue, conditioned on the initial length of the transmit buffer [10]. Having this distribution, in turn, permits the computation of several performance criteria such as mean first-passage time to an empty queue and the probability of violating a delay threshold. Previously proposed analysis techniques to derive distributions of waiting times can account for channel correlation across codewords and they work well for small buffer sizes, which are typical of communication systems subject to stringent delay restrictions. However, they become cumbersome for large initial buffers, which are more representative of delay-aware applications with softer constraints.

In such scenarios, analyzing the large deviations governing the evolution of the system offers a promising new direction to derive meaningful guidelines for resource allocation and the selection of system parameters. Specifically, the concentration of empirical measures can be used to gracefully adjust delay sensitivity to the needs of a real-time data flow by selecting the deviation threshold, i.e., the argument of the rate function [11]. Once a threshold is set, system parameters can be optimized according to this objective function and the resulting performance can be predicted accurately.

The remainder of this article is devoted to finding appropriate large deviation principles for the problem at hand. A mathematical model for the communication system under

This material is based upon work supported, in part, by the National Science Foundation under Grant No. 0747363 and Grant No. 0830696. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

consideration is described briefly in the next section. In Section III, the evolution of the transmit buffer is derived in terms of the properties of the physical layer. Based on the proposed model, two types of aberrations are examined: deviations in average service opportunities and mean packet transmission time. Numerical results in the context of practical implementations are presented in Section IV. Conclusions, design guidelines and possible extensions are discussed in the last section of the paper.

## II. System Model

In this study, information is assumed to flow from a source to a destination over an unreliable communication link. To protect a message against erasures, its content is fragmented into segments of size $K$ and redundancy is added to every data block using random codes. Every encoded packet is sent to the destination over the wireless channel. Contingent on the channel realization, a segment is either successfully decoded and its reception acknowledged or, upon decoding failure, immediate retransmission of the segment is triggered. This transmission paradigm is common to many contemporary communication systems. A distinguishing aspect of our line of work lies in the rigorous treatment of the system evolution at the link layer based on a channel abstraction at the symbol level that incorporates correlation over time.

### A. Erasure Channel with Memory

The connection between the two wireless devices is modeled as a finite-state erasure channel. Again, we stress that such models have been introduced and used in the literature in the past [7], [8], [12], [10]. Thus, the discussion of this abstraction for channels with memory is kept to a minimum, with the understanding that detailed explanations have already been published. Still, for the sake of completeness, the required notation is reviewed.

The communication channel operates in one of finitely many states. These states are collectively denoted by $\mathcal{C} = \{c_1, \ldots, c_k\}$. Channel transitions over time form a Markov chain with stochastic matrix

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kk} \end{bmatrix}. \tag{1}$$

The entry $b_{ij}$ represents the probability of going from state $c_i$ to state $c_j$ in one channel use. When in state $c_i$, the probability that a transmitted symbol gets erased is equal to $\varepsilon_i$; it is therefore received faithfully with probability $1 - \varepsilon_i$. For notational convenience, we assume that state indices are assigned such that $\varepsilon_i \geq \varepsilon_j$ whenever $i < j$. We also assume that $\mathbf{B}$ is irreducible. The state of the channel at time $n$ is a random variable, which we denote by $C_n$, and its evolution over time forms a stochastic process $\{C_n\}$.

### B. Coding Strategy

Data segments are encoded into codewords of length $N$ using random codebooks; each codebook is determined as follows. For every transmission attempt, a parity-check matrix $\mathbf{H}_s$ of size $(N-K) \times N$ is chosen randomly. Each entry in this matrix is either zero or one with equal probability, and the entries are selected independently of one another. The set of admissible codewords corresponds to the nullspace of $\mathbf{H}_s$. Information segments are then assigned to codewords, and maximum-likelihood decoding is performed at the destination. This coding scheme is known to perform well in the large block-length regime and, given $e$ erasures within a block, the probability of decoding failure can be written as [13, p. 164]

$$P_{\mathrm{f}}(N-K, e) = 1 - \prod_{i=0}^{e-1} \left(1 - 2^{i-(N-K)}\right). \tag{2}$$

It is immediately apparent from this expression that the distribution of $E$, the number of erasures within a block, plays a major role in determining overall system performance. In particular, the probability of decoding failure conditioned on initial channel state $c_i$ is equal to

$$\sum_{c_j \in \mathcal{C}} \sum_{e=0}^{N} P_{\mathrm{f}}(N-K, e) \Pr\left(E = e, C_{N+1} = c_j | C_1 = c_i\right).$$

From this, we see that $\Pr\left(E = e, C_{N+1} = c_j | C_1 = c_i\right)$ is a quantity of interest in understanding failure events. Luckily, there are efficient means to compute the conditional distribution of the number of erasures within a block for finite-state channels with memory [14], [15]. For instance, let $\mathbf{B}_x$ be the matrix of polynomials defined entry-wise by

$$[\mathbf{B}_x]_{ij} = b_{ij}(1 - \varepsilon_i + \varepsilon_i x) \quad 1 \leq i, j \leq k.$$

Then, for $e \in \mathbb{N}_0$ and $c_i, c_j \in \mathcal{C}$,

$$\Pr(E = e, C_{N+1} = c_j | C_1 = c_i) = [\![x^e]\!] \left[\mathbf{B}_x^N\right]_{i,j}, \tag{3}$$

where $[\![x^e]\!]$ represents the operator that maps a polynomial in $x$ to the coefficient of $x^e$. In some sense, computing $\mathbf{B}_x^N$ is simply an efficient way to concurrently sum over all relevant outcomes of the Markov process. While there are other means to obtain the desired conditional probabilities (e.g., nested sums), this technique is very efficient and it will be leveraged again when discussing the evolution of the queue in the next section.

## III. Performance Characterization

To apply large deviations techniques to the problem at hand, we must first understand the evolution of the packetized system over time. As mentioned above, segments of $K$ information bits are encoded into packets of length $N$, which are then sent over the physical channel. The transmission of a codeword thus necessitates $N$ consecutive uses of the channel. A service opportunity occurs every time the random code and channel realization jointly permit reliable decoding. This is possible if and only if the submatrix of $\mathbf{H}_s$ formed by selecting the

$e$ erased columns has rank $e$ [13, p. 73]. Let $D_s$ indicate the potential of a successful decoding event. That is, $D_s = 1$ when a message can be (or would be) decoded faithfully; and $D_s = 0$ otherwise. Hence, the stochastic process $\{D_s\}$ identifies time instants at which blocks of information can potentially be transferred successfully to the destination.

We consider the random vector $V_s = (C_{(s+1)N+1}, D_s)$, composed of the state of the erasure channel at the onset of block $s+1$, together with the indicator of a service opportunity during block $s$. The stochastic process $\{V_s\}$ is a Markov chain and its transition probabilities can be derived from the properties of the underlying erasure channel and the coding scheme introduced above. More specifically, they can be obtained by conditioning on the number of erasures within a block,

$$\begin{aligned}
&\Pr(V_{s+1} = (c_j, d_{s+1}) | V_s = (c_i, d_s)) \\
&\quad = \sum_{e \in \mathbb{N}_0} \Pr(V_{s+1} = (c_j, d_{s+1}), E = e | V_s = (c_i, d_s)) \\
&\quad = \sum_{e \in \mathbb{N}_0} \Pr\left(E = e, C_{(s+2)N+1} = c_j | C_{(s+1)N+1} = c_i\right) \times \\
&\qquad \Pr\left(D_{s+1} = d_{s+1} | E = e\right) \qquad (4)
\end{aligned}$$

where $c_i, c_j \in \mathcal{C}$ and $d_s, d_{s+1} \in \{0, 1\}$. The first component of the summand is the conditional distribution of the number of erasures within a block, as defined in (3). The second part is governed by (2) with

$$\begin{aligned}
\Pr(D_{s+1} = 1 | E = e) &= 1 - P_{\mathrm{f}}(N-K, e) \\
\Pr(D_{s+1} = 0 | E = e) &= P_{\mathrm{f}}(N-K, e).
\end{aligned} \qquad (5)$$

We emphasize that decoding events may only take place upon completion of codeword transmissions, i.e., every $N$ channel uses. This explains the discrepancy between the indexing of service opportunities and channel states.

The service opportunities over time provide a first level of insight into communication links. Another significant consideration is the evolution of the transmit queue at the source. In this article, we are interested in the first-passage time to an empty buffer and, to capture this quantity adequately, we need to characterize the evolution of the queue. Suppose that the transmit buffer at the source initially contains $\ell$ information bits that need to be sent to the destination over the wireless link. These bits are partitioned into $m = \lceil \ell / K \rceil$ data segments, which are successively encoded and transferred over the channel. Each segment remains in the transmit buffer until the destination reliably acquires its content. Once accurate reception is properly acknowledged by the destination, the matching information bits are discarded at the source, and transmission of the next data segment begins.

The size of the transmit queue at the onset of block $s$ is represented by $Q_s$. The state of the erasure channel at that same time instant is $C_{sN+1}$. The random vector $U_s = (C_{sN+1}, Q_s)$, composed of the channel state and queue length, contains all the relevant information to track the evolution of the system. The process $\{U_s\}$ provides the foundation of our second large deviations analysis. Since there are no arrivals in our framework, the evolution of the queue is governed by

$$Q_{s+1} = (Q_s - D_s)^+ . \qquad (6)$$

The augmented stochastic process $\{U_s\}$ also forms a Markov chain [10]. It is straightforward to relate the transition probabilities of $Q_s$ to those of $D_s$, as specified in (5). The only two possibilities for the evolution of a non-empty queue are to remain at a same level or to go down one level. A graphical interpretation of the state diagram for the queued system is depicted in Fig. 1. The $q$th level of the diagram refers to the set $\{(c, q) | c \in \mathcal{C}\}$.

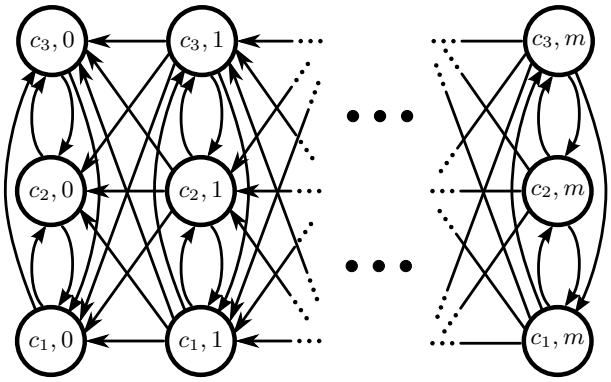


Fig. 1. This figure shows transitions in queue length and channel state for a communication system that sends data over a three-state erasure channel. The state space corresponds to the augmented process $U_s = (C_{(s-1)N+1}, Q_s)$ and possible transitions are depicted by arrows, expect self-transitions which are intentionally omitted.

For tractability, we introduce a uniform notation for the transition probabilities of our two augmented Markov processes,

$$\begin{aligned}
\kappa_{ij} &= \Pr(V_{s+1} = (c_j, 0) | V_s = (c_i, d)) \\
&= \Pr(U_{s+1} = (c_j, q) | U_s = (c_i, q)) \\
\mu_{ij} &= \Pr(V_{s+1} = (c_j, 1) | V_s = (c_i, d)) \\
&= \Pr(U_{s+1} = (c_j, q-1) | U_s = (c_i, q))
\end{aligned} \qquad (7)$$

where $q \in \mathbb{N}$, $c_i, c_j \in \mathcal{C}$ and $d \in \{0, 1\}$. This notation also emphasizes the close relation between $\{V_s\}$ and $\{U_s\}$.

### A. Empirical Mean Service

In this section, we focus on the large deviations of the empirical mean service

$$Z_s = \frac{1}{s} \sum_{t=1}^{s} D_t .$$

We note that $\{D_s\}$ is not a Markov process. However, $D_s$ is a (trivial) deterministic function of $V_s = (C_{(s+1)N+1}, D_s)$. Since $\{V_s\}$ is a Markov process, we can apply general results on the large deviation principle of additive functionals of Markov chains. To leverage these results, we first impose an ordering on the state space $\mathcal{V} = \mathcal{C} \times \{0, 1\}$. Recall that $|\mathcal{C}| = k$; a natural ordering for this state space is to associate integer $v = (dk + i)$ with state $(c_i, d)$. Using this ordering,

the transition probability matrix $\mathbf{\Pi}$ for the augmented process $\{V_s\}$ is defined by

$$[\mathbf{\Pi}]_{v_1,v_2} = \pi(v_1, v_2), \quad v_1, v_2 \in \{1, \ldots, 2k\}$$

where $\pi(v_1, v_2)$ is the probability of jumping to state $v_2$, conditioned on starting from $v_1$. Since $\mathbf{\Pi}$ is irreducible, the following theorem applies.

*Theorem 1 (Dembo & Zeitouni [11]):* Let $\{V_s\}$ be a finite state Markov chain possessing an irreducible transition matrix $\mathbf{\Pi}$. For every $x \in \mathbb{R}$, define

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \log \varrho\,(\mathbf{\Pi}_\lambda)\} \tag{8}$$

where $\mathbf{\Pi}_\lambda$ is a nonnegative matrix whose elements are

$$\pi_\lambda\,(v_1, v_2) = \pi(v_1, v_2) e^{\lambda d_2} \quad v_1, v_2 \in \{1, \ldots, 2k\}$$

and $\varrho(\mathbf{\Pi}_\lambda)$ denotes the spectral radius of the matrix argument. Then, the empirical mean $Z_s$ satisfies the large deviation principle with the convex good rate function $I(\cdot)$. Explicitly, for any set $\Gamma \subseteq \mathbb{R}$, and any initial state $v \in \mathcal{V}$,

$$\begin{aligned} -\inf_{x \in \Gamma^o} I(x) &\le \liminf_{s\to\infty} \frac{1}{s} \log P_v^\pi(Z_s \in \Gamma) \\ &\le \limsup_{s\to\infty} \frac{1}{s} \log P_v^\pi(Z_s \in \Gamma) \le -\inf_{x \in \bar{\Gamma}} I(x) \end{aligned}$$

where $P_v^\pi$ denotes the Markov probability measure induced by transition probability $\mathbf{\Pi}$ and initial state $v \in \mathcal{V}$, i.e.,

$$P_v^\pi(V_1 = v_1, \ldots, V_s = v_s) = \pi(v, v_1) \prod_{t=1}^{s-1} \pi(v_t, v_{t+1}).$$

Expressions for the transition probabilities used in this theorem appear in (7). We note that

$$\begin{aligned} &\Pr(V_{s+1} = (c_j, d_2) | V_s = (c_i, d_1)) \\ &= \Pr(V_{s+1} = (c_j, d_2) | C_{(s+1)N+1} = c_i); \end{aligned}$$

this induces a repetitive structure in matrix $\mathbf{\Pi}$. The nonnegative matrix $\mathbf{\Pi}_\lambda$ associated with every $\lambda \in \mathbb{R}$ can then be given explicitly as

$$\mathbf{\Pi}_\lambda = \begin{bmatrix} \kappa_{11} & \cdots & \kappa_{1k} & \mu_{11}e^\lambda & \cdots & \mu_{1k}e^\lambda \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} & \mu_{k1}e^\lambda & \cdots & \mu_{kk}e^\lambda \\ \kappa_{11} & \cdots & \kappa_{1k} & \mu_{11}e^\lambda & \cdots & \mu_{1k}e^\lambda \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} & \mu_{k1}e^\lambda & \cdots & \mu_{kk}e^\lambda \end{bmatrix}. \tag{9}$$

We can rewrite $\mathbf{\Pi}_\lambda$ by taking advantage of its block structure,

$$\mathbf{\Pi}_\lambda = \begin{bmatrix} \mathbf{K} & \mathbf{M}e^\lambda \\ \mathbf{K} & \mathbf{M}e^\lambda \end{bmatrix}$$

where the submatrices introduced above are defined as

$$\mathbf{K} = \begin{bmatrix} \kappa_{11} & \cdots & \kappa_{1k} \\ \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1k} \\ \vdots & \ddots & \vdots \\ \mu_{k1} & \cdots & \mu_{kk} \end{bmatrix}.$$

The pertinent eigenvalues are the roots of the characteristic polynomial of $\mathbf{\Pi}_\lambda$. Using properties of matrix determinant and the commutative properties of some of the blocks, we can express this polynomial as

$$\begin{aligned} \det(\gamma\mathbf{I} - \mathbf{\Pi}_\lambda) &= \det\left((\gamma\mathbf{I} - \mathbf{M}e^\lambda)(\gamma\mathbf{I} - \mathbf{K}) - \mathbf{M}\mathbf{K}e^\lambda\right) \\ &= \det\left((\gamma\mathbf{I} - \mathbf{K})(\gamma\mathbf{I} - \mathbf{M}e^\lambda) - \mathbf{K}\mathbf{M}e^\lambda\right). \end{aligned}$$

Collectively, Theorem 1 and the matrix defined in (9) provide an algorithmic work-flow for the computation of the good rate function associated with the empirical means $\{Z_s\}$. We follow this discussion with a simple example based on a two-state channel with memory.

*Example 1:* Consider a Gilbert-Elliott erasure channel [7], [8], with $\mathcal{C} = \{c_1, c_2\}$. An advantage in studying this rudimentary model is that it admits a simple, closed-form characterization. The dimension of the state space in this case is $|\mathcal{V}| = 4$. Using the commutative block structure discussed above, we can write the determinant of $(\gamma\mathbf{I} - \mathbf{\Pi}_\lambda)$ as

$$\begin{aligned} &\det(\gamma\mathbf{I} - \mathbf{\Pi}_\lambda) = \det\left(\gamma^2\mathbf{I} - \gamma\mathbf{K} - \gamma\mathbf{M}e^\lambda\right) \\ &= \gamma^2 \det\left(\begin{bmatrix} \gamma - \kappa_{11} - \mu_{11}e^\lambda & -\kappa_{12} - \mu_{12}e^\lambda \\ -\kappa_{21} - \mu_{21}e^\lambda & \gamma - \kappa_{22} - \mu_{22}e^\lambda \end{bmatrix}\right). \end{aligned}$$

By inspection, we see that the spectral radius of $\mathbf{\Pi}_\lambda$ is the largest root of the quadratic equation

$$\begin{aligned} &\gamma^2 - \gamma(\kappa_{11} + \kappa_{22} + (\mu_{11} + \mu_{22})e^\lambda) \\ &+ (\kappa_{11} + \mu_{11}e^\lambda)(\kappa_{22} + \mu_{22}e^\lambda) \\ &- (\kappa_{12} + \mu_{12}e^\lambda)(\kappa_{21} + \mu_{21}e^\lambda) = 0. \end{aligned}$$

We will revisit this example in Section IV.

### *B. First-Passage Time*

We turn to the second type of aberrations we wish to study: deviations in the normalized first-passage time to an empty queue. Again, suppose that the transmit buffer contains exactly $m$ segments at the onset of the communication process. In this case, the data transfer terminates after $m$ successful decoding events. To capture the completion time of this process, we first explore properties of the sojourn time at a given level.

The evolution of the queue is regulated by the current state of the buffer and the occurrence of a service opportunity, as seen in (6). Two collections of random variables that are of fundamental importance in our queueing analysis are the hitting times to the $q$th level of the chain

$$H_q = \inf\{s \ge 0 | Q_s = q\}, \tag{10}$$

and the sojourn time at the $q$th level of the chain

$$T_q = H_{q-1} - H_q.$$

The variable $H_q$ represents the moment at which the Markov chain $\{U_s\}$ first enters its $q$th level. Moreover, $T_q$ is the total time $\{U_s\}$ spends at level $q$ before moving to a lower level. Of course, these quantities only make sense for values of $q$ that are less than or equal to the initial size of the queue.

With this notation, we can write the hitting time $H_0$ as a sum of random variables

$$H_0 = \sum_{q=1}^{m} T_q.$$

Looking at the Markov property of the finite-state channel, we gather that the sojourn times at the various levels are conditionally independent, given the state of the channel at time instants $\{H_q\}_{q=0}^{m}$. It is thus possible to derive the distribution of $H_0$ using polynomials and generating matrices [10].

We first consider a reduced problem. Suppose that the augmented Markov chain starts at level $q > 0$ with channel state distribution $\pi_q$. We wish to compute the joint probability distribution of $T_q$ and $C_{NH_{q-1}+1}$. In other words, we want to characterize the probability distribution of the sojourn time at level $q$ together with the probability that the channel enters state $c_j$ as the chain reaches the subsequent lower level. Assuming for now that states at level $q-1$ are all absorbing, we can write the transition probability of this degenerate subsystem as

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{M} & \mathbf{K} \end{bmatrix}. \tag{11}$$

Using the Neumann expansion, we get

$$\begin{aligned} \pi_{q-1} &= \begin{bmatrix} \mathbf{0} & \pi_q \end{bmatrix} \left( \lim_{t\to\infty} \mathbf{P}^t \right) \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \pi_q \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I}-\mathbf{K})^{-1}\mathbf{M} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \\ &= \pi_q (\mathbf{I}-\mathbf{K})^{-1}\mathbf{M}. \end{aligned}$$

where $\pi_{q-1}(j) = \Pr(C_{NH_{q-1}+1} = c_j)$. This standard technique provides the distribution of the channel state as the chain reaches the lower level. To account for the sojourn time, we resort to abstract polynomials. Let generating matrix $\mathbf{G}_T(z)$ be defined component-wise by

$$[\mathbf{G}_T(z)]_{ij} = \mathbf{E}\left[ z^{T_q} \mathbf{1}_{\{C_{NH_{q-1}+1}=c_j\}} | C_{NH_q+1} = c_i \right] \tag{12}$$

where $\mathbf{1}_{\{\cdot\}}$ is the set indicator function. We point out that this expression remains unaltered for $0 < q \leq m$; this hints at the regular structure of this queueing problem.

*Lemma 1 (Chamberland et al. [10]):* For the reduced subsystem associated with (11), the generating matrix $\mathbf{G}_T(z)$ is equal to

$$\mathbf{G}_T(z) = \left( \sum_{t=0}^{\infty} \mathbf{K}^t z^t \right) \mathbf{M}z = (\mathbf{I}-\mathbf{K}z)^{-1}\mathbf{M}z. \tag{13}$$

This result can be obtained by mimicking the derivation above using the generalized transition matrix over real polynomials,

$$\mathbf{P}_z = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{M}z & \mathbf{K}z \end{bmatrix}.$$

With this construction, we get

$$\Pr\left(T = t, C_{NH_{q-1}+1} = c_j\right) = [\![ z^t ]\!] \left[\pi_q \mathbf{G}_T(z)\right]_j.$$

Applying the principle of mathematical induction to this result, we can derive the following proposition.

*Proposition 2 (Chamberland et al. [10]):* The ordinary generating function of $H_0$, the first-passage time to an empty queue, is given by

$$G_{H_0}(z) = \pi_0 \left(\mathbf{G}_T(z)\right)^m \mathbf{1}, \tag{14}$$

where $\pi_0$ is the distribution of the channel states at time zero and $\mathbf{1}$ is a column vector of all ones.

To differentiate among possible initial conditions, we can write the first-passage time to an empty queue as $H_0^{(m)}$ where the superscript represents the number of segments in the queue at time zero. We employ a similar strategy to distinguish generating functions. As mentioned above, we are interested in the large deviations associated with the sequence of random variables specified by

$$Y_m = \frac{1}{m} H_0^{(m)} = \frac{1}{m} \sum_{q=1}^{m} T_q \quad m = 1, 2, \ldots$$

The logarithmic moment generating function of $Y_m$ is

$$\begin{aligned} \Lambda_m(\lambda) &= \log \mathbf{E}\left[e^{\lambda Y_m}\right] = \log \mathbf{E}\left[e^{\lambda H_0^{(m)}/m}\right] \\ &= \log G_{H_0}^{(m)}\left(e^{\lambda/m}\right). \end{aligned}$$

The existence of limits of properly scaled logarithmic moment generating functions suggests that $\{Y_m\}$ may satisfy a large deviation principle. In particular, consider the following asymptotic regime

$$\begin{aligned} \Lambda(\lambda) &= \lim_{m\to\infty} \frac{1}{m} \Lambda_m(m\lambda) = \lim_{m\to\infty} \frac{1}{m} \log G_{H_0}^{(m)}\left(e^{\lambda}\right) \\ &= \lim_{m\to\infty} \frac{1}{m} \log \left( \pi_0 \left( \mathbf{G}_T\left(e^{\lambda}\right) \right)^m \mathbf{1} \right). \end{aligned} \tag{15}$$

A few observations concerning $\Lambda(\lambda)$ are in order. We remark that, for any $z = e^{\lambda}$,

$$\mathbf{G}_T\left(e^{\lambda}\right) = \left( \sum_{t=0}^{\infty} \mathbf{K}^t e^{t\lambda} \right) \mathbf{M}e^{\lambda}$$

is a positive matrix over the extended real numbers. This matrix is real if and only if $e^{\lambda} < 1/\varrho(\mathbf{K})$. Furthermore, whenever this condition holds, $\mathbf{G}_T\left(e^{\lambda}\right)$ is irreducible and the Perron-Frobenius theorem applies [16], [11]. This leads to the following proposition.

*Proposition 3:* For each $\lambda \in \mathbb{R}$, the moment generating function defined in (15) exists as an extended real number with

$$\Lambda(\lambda) = \begin{cases} \varrho\left( \left(\mathbf{I} - \mathbf{K}e^{\lambda}\right)^{-1} \mathbf{M}e^{\lambda} \right) & \lambda < -\log \varrho(\mathbf{K}) \\ \infty & \text{otherwise.} \end{cases}$$

Again, $\varrho(\cdot)$ denotes the spectral radius of its matrix argument.

*Proof:* Suppose $\lambda < -\log \varrho(\mathbf{K})$. Then, the spectral radius $\varrho\left(\mathbf{K}e^{\lambda}\right)$ is strictly less than one and $(\mathbf{I} - \mathbf{K}e^{\lambda})$ is invertible. This implies that

$$\mathbf{G}_T\left(e^{\lambda}\right) = \left( \sum_{t=0}^{\infty} \mathbf{K}^t e^{t\lambda} \right) \mathbf{M}e^{\lambda} = \left(\mathbf{I} - \mathbf{K}e^{\lambda}\right)^{-1} \mathbf{M}e^{\lambda}$$

is well-defined over the real numbers. Let $\mathbf{w}_\lambda$ be the right eigenvector of $\mathbf{G}_T\left(e^\lambda\right)$ that corresponds to the Perron-Frobenius eigenvalue. Note that all the components of $\mathbf{w}_\lambda$ are positive and hence $\langle \mathbf{1}|\mathbf{w}_\lambda\rangle > 0$. As such, we have

$$\begin{aligned}
&\pi_0\left(\mathbf{G}_T\left(e^\lambda\right)\right)^m \mathbf{1} \\
&= \pi_0\left(\mathbf{G}_T\left(e^\lambda\right)\right)^m \left(\frac{\langle \mathbf{1}|\mathbf{w}_\lambda\rangle}{\|\mathbf{w}_\lambda\|^2}\mathbf{w}_\lambda + \left(\mathbf{1} - \frac{\langle \mathbf{1}|\mathbf{w}_\lambda\rangle}{\|\mathbf{w}_\lambda\|^2}\mathbf{w}_\lambda\right)\right) \\
&= \left(\varrho\left(\left(\mathbf{I}-\mathbf{K}e^\lambda\right)^{-1}\mathbf{M}e^\lambda\right)\right)^m \frac{\langle \mathbf{1}|\mathbf{w}_\lambda\rangle}{\|\mathbf{w}_\lambda\|^2}\pi_0\mathbf{w}_\lambda \\
&+ o\left(\left(\varrho\left(\left(\mathbf{I}-\mathbf{K}e^\lambda\right)^{-1}\mathbf{M}e^\lambda\right)\right)^m\right),
\end{aligned}$$

This immediately implies that

$$\Lambda(\lambda) = \varrho\left(\left(\mathbf{I}-\mathbf{K}e^\lambda\right)^{-1}\mathbf{M}e^\lambda\right)$$

for $\lambda < -\log\varrho(\mathbf{K})$.

On the other hand, assume that $\lambda \geq -\log\varrho(\mathbf{K})$. We denote the right eigenvector of $\mathbf{K}$ that corresponds to its Perron-Frobenius eigenvalue by $\mathbf{w}$. We stress that $\mathbf{M}e^\lambda\mathbf{1}$ is a vector with positive components and hence

$$\left\langle \mathbf{M}e^\lambda\mathbf{1}|\mathbf{w}\right\rangle > 0.$$

Thus, the sequence of vectors defined by

$$\left(\sum_{s=0}^{t}\mathbf{K}^s e^{s\lambda}\right)\mathbf{M}e^\lambda\mathbf{1}$$

diverges entry-wise to infinity and, therefore, $\Lambda(\lambda)=\infty$. This completes the proof of Proposition 3. ∎

Using matrix norms, it can be shown that $\mathbf{G}_T\left(e^\lambda\right)$ is differentiable entry-wise over the interval $\lambda < -\log\varrho(\mathbf{K})$. Since $\Lambda(\lambda)$ is an isolated root of the characteristic function of matrix $\mathbf{G}_T\left(e^\lambda\right)$, we deduce that it is positive, finite and differentiable with respect to $\lambda$ [17]. Consequently, $\Lambda(\lambda)$ is essentially smooth and the Gärtner-Ellis theorem applies [11]. This leads to the following proposition.

*Theorem 4:* Let $\left\{Y_m = \frac{1}{m}\sum_{q=1}^{m} T_q\right\}$ be the empirical mean sojourn time per level. For every $x\in\mathbb{R}$, consider the Fenchel-Legendre transform

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}}\left\{\lambda x - \log\varrho\left(\mathbf{G}_T\left(e^\lambda\right)\right)\right\}. \tag{16}$$

The empirical mean $Y_m$ satisfies the large deviation principle with the convex, good rate function $\Lambda^*(\cdot)$. That is, for any set $\Gamma\subseteq\mathbb{R}$ and any initial state $c\in\mathcal{C}$,

$$\begin{aligned}
-\inf_{x\in\Gamma^o}\Lambda^*(x) &\leq \liminf_{m\to\infty}\frac{1}{m}\log\Pr(Y_m\in\Gamma) \\
&\leq \limsup_{m\to\infty}\frac{1}{m}\log\Pr(Y_m\in\Gamma) \leq -\inf_{x\in\bar{\Gamma}}\Lambda^*(x).
\end{aligned}$$

*Example 2:* For the Gilbert-Elliott channel introduced in Example 1, it is possible to obtain a closed-form expression for the spectral radius of $\mathbf{G}_T\left(e^\lambda\right)$. Specifically, we can write the characteristic polynomial of $\mathbf{G}_T\left(e^\lambda\right)$ as

$$\begin{aligned}
\det\left(\gamma\mathbf{I}-\mathbf{G}_T\left(e^\lambda\right)\right) &= \det\left(\gamma\mathbf{I}-\left(\mathbf{I}-\mathbf{K}e^\lambda\right)^{-1}\mathbf{M}e^\lambda\right) \\
&= \frac{\det\left(\gamma\mathbf{I}-\gamma\mathbf{K}e^\lambda-\mathbf{M}e^\lambda\right)}{\det\left(\mathbf{I}-\mathbf{K}e^\lambda\right)}.
\end{aligned}$$

We note that the numerator is a quadratic equation and the denominator is a constant. It is therefore possible to find parametric expressions for the two roots of $\det\left(\gamma\mathbf{I}-\mathbf{G}_T\left(e^\lambda\right)\right)$. Taking the maximum of the absolute values of these two roots yields an explicit, albeit convoluted, expression for the spectral radius of $\mathbf{G}_T\left(e^\lambda\right)$. In any case, $\Lambda^*(\cdot)$ can be evaluated efficiently using numerical methods.

Having found expressions for large deviations on the average service opportunity and mean sojourn time, we turn to a numerical study that will produce helpful insights and meaningful guidelines.

## IV. Numerical Results

Above, we have introduced a mathematical machinery that works well for specific block-lengths and code rates. In this section, we seek to compare performance as functions of these design parameters. In particular, we want to compare systems that employ different values for $K$ and $N$. Consequently, we need to introduce a commonality for the scaling of competing implementations. A proper scaling for a fair comparison of average services can be defined in terms of decoded bits per channel use,

$$\tilde{Z}_n = \frac{1}{n}\sum_{t=1}^{\lfloor n/N\rfloor} KD_t.$$

This leads to the following asymptotic regime

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\log\Pr\left(\tilde{Z}_n<\tau\right) \\
&= \frac{1}{N}\lim_{n\to\infty}\frac{1}{\lfloor n/N\rfloor}\log\Pr\left(\frac{1}{\lfloor n/N\rfloor}\sum_{t=1}^{\lfloor n/N\rfloor}D_t<\frac{N}{K}\tau\right) \\
&= \frac{1}{N}\lim_{s\to\infty}\frac{1}{s}\log\Pr\left(\frac{1}{s}\sum_{t=1}^{s}D_t<\frac{N}{K}\tau\right) = -\frac{1}{N}I\left(\frac{N}{K}\tau\right)
\end{aligned}$$

where $\tau < \mathrm{E}\left[\tilde{Z}_\infty\right]$. Similarly, to account for discrepancies in design parameters, the empirical mean sojourn time can be expressed in terms of channel uses per information bit,

$$\tilde{Y}_\ell = \frac{1}{\ell}NH_0^{\lceil \ell/K\rceil}.$$

The corresponding asymptotic regime then becomes

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{\ell}\log\Pr\left(\tilde{Y}_\ell>\kappa\right) \\
&= \frac{1}{K}\lim_{\ell\to\infty}\frac{1}{\lceil \ell/K\rceil}\log\Pr\left(\frac{1}{\lceil \ell/K\rceil}H_0^{(\lceil \ell/K\rceil)}>\frac{K}{N}\kappa\right) \\
&= \frac{1}{K}\lim_{m\to\infty}\frac{1}{m}\log\Pr\left(\frac{1}{m}H_0^{(m)}>\frac{K}{N}\kappa\right) = -\frac{1}{K}\Lambda^*\left(\frac{K}{N}\kappa\right)
\end{aligned}$$

where $\kappa > \mathrm{E}\left[\tilde{Y}_\infty\right]$. With these important adjustments to the rate functions, we are now ready to compare the performance of different systems. The fundamental tradeoff underlying code-rate selection can be summarized as follows. Low-rate codes offer better protection against erasures and are therefore more likely to be successfully decoded. However, high-rate codes reveal more information bits upon successful packet

decoding. Reconciling these two competing considerations is the key to maximizing overall performance.

Varying block-length, as opposed to code rate, leads to more subtle considerations. The length of a codeword, which we labeled $N$, implicitly dictates the amount of feedback needed for acknowledgments, with short block-lengths requiring more feedback. A careful formulation that provides flexibility in choosing $N$ must account for the cost of feedback. For the numerical examples presented below, we circumvent these difficulties and assume that $N$ is fixed. The optimization is carried solely over $K$, leading to a code rate of $K/N$.

An important application of wireless systems is voice communication, which is quite sensitive to delay. In GSM, each speech frame of length 20 ms is encoded into a bit-stream of length 228. The underlying physical layer has the ability to transmit one bit every 40 $\mu$s. If we approximate the maximum delay tolerance for voice traffic over this wireless connection to be 40 ms [2, p. 70], then this requires 228 bits to be transmitted within 1000 channel uses. This constraint, in turn, necessitates a nominal rate of approximately 0.23 bits per channel use for link reliability. Thus, for the purpose of this example, we can set a value of 1000 for $n$ and a deviation threshold of 0.23 for $\tau$, and minimize the probability with which the service rate drops below this threshold.

The following parameters are selected for our numerical study. The channel is a two-state Gilbert-Elliott erasure channel with $\varepsilon_1 = 1$ and $\varepsilon_2 = 0$ and the average erasure rate is taken to be

$$\frac{b_{21}}{b_{12}+b_{21}} = 0.2.$$

The channel memory for this model can be quantified through the second largest eigenvalue of the channel transition probability matrix $\mathbf{B}$, and is set at $1 - b_{12} - b_{21} = 0.9$. The block length is $N = 114$ and $K$ is the parameter to be optimized.

The rate function governing the large deviations for the empirical mean service appears in Fig. 2. The threshold $\tau$ represents a minimum target requirement on the number of information bits per channel use that can be successfully decoded at the destination, in the asymptotic regime. The optimal value of $K$ as a function of target threshold $\tau$ corresponds to the apex of each curve,

$$K_Z^*(\tau) = \arg\max_K \frac{1}{N} I\left(\frac{N}{K}\tau\right).$$

At low code-rates, the maximum achievable throughput is less than the service requirement and hence the rate function governing large deviations is zero. Moreover, at high code-rates, performance is limited by the rise in the probability of decoding failure. The system needs to balance the probability of a successful decoding event with its reward in terms of information bits. It is interesting to note how conservative the optimal code-rate becomes when the target service requirement is reduced.

Rate functions that characterize large deviations in the mean sojourn times are plotted in Fig. 3. Variable $\kappa$ represents an upper threshold requirement on the average number of

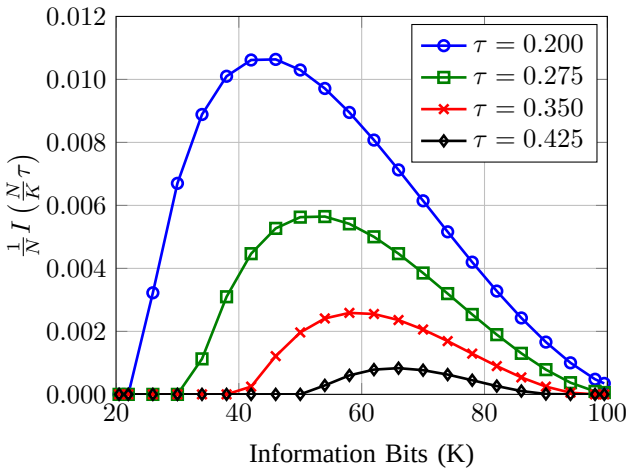


Fig. 2. This figure plots good rate functions governing large deviations in the empirical mean service as functions of $K$, the number of information bits per codeword. Given throughput threshold $\tau$, the optimal value of $K$ is the argument corresponding to the apex of the function.

channel uses employed to transmit one information bit. The behavior of the system in terms of average sojourn time is closely related to the empirical mean service, holding an approximately reciprocal relation. In this case, the optimal value of $K$ becomes

$$K_Y^*(\kappa) = \arg\max_K \frac{1}{K}\Lambda^*\left(\frac{K}{N}\kappa\right).$$

Interestingly, when $\kappa = 1/\tau$, the optimal code-rates are equal, namely $K_Z^*(\tau) = K_Y^*(\kappa)$. This agreement between optimal code rates appears robust to the choice of channel parameters.

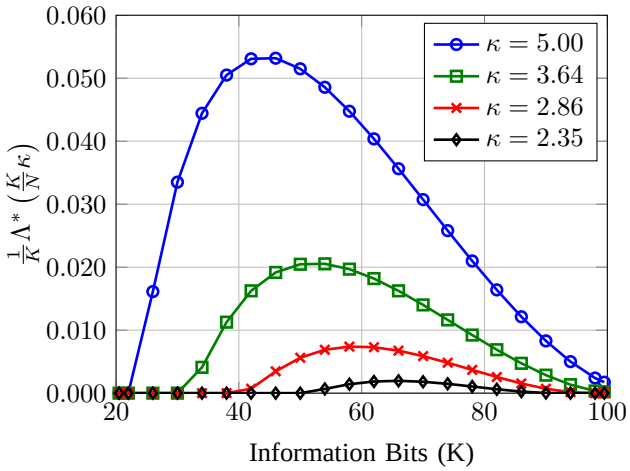


Fig. 3. This figure shows good rate functions governing large deviations in the mean sojourn time as functions of $K$. The optimum code-rate depends heavily on the deviation threshold of the mean sojourn time.

## V. Guidelines and Possible Extensions

This article provides a methodology for selecting code rate based on service or delay requirements. For both average

service and mean sojourn time, it seems that the optimal operating point of a system in terms of code rate selection depends heavily on the needs of the underlying traffic. In particular, delay-adverse applications may perform better with coarse quantization and low-rate codes. On the other hand, a delay tolerant application may be able to get a much higher rate out of a same physical channel. This phenomenon is closely related to the concept of effective capacity.

An interesting future research topic is to have traffic-aware devices adapt to changing traffic conditions and multi-tasking devices with heterogeneous flows. One possible approach would be to pose this problem in a utility maximization framework, using different reward functions for the possible applications. Another interesting aspect of this research is that it offers tools to quantify the quality of various channels, as seen from an application point of view. A low-rate cellular connection with a specular component may be well-suited for voice, owing to its low variability. On the other hand, a slow-varying Rayleigh channel may be preferable for web traffic and bandwidth-hungry applications. In systems with multiple access points, this suggests looking for channels best suited for their current traffic profiles.

## REFERENCES


[1] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, 1994.

[2] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. New York, NY, USA: Cambridge University Press, 2005.

[3] Y. Polyanskiy, H. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, April 2011.

[4] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 59–68, September 2004.

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

[6] S. Shakkottai, "Effective capacity and QoS for wireless scheduling," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 749–761, April 2008.

[7] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J*, vol. 39, no. 5, pp. 1253–1265, 1960.

[8] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J*, vol. 42, no. 5, pp. 1977–1997, 1963.

[9] P. Parag, J.-F. Chamberland, H. Pfister, and K. R. Narayanan, "On the queueing behavior of random codes over a Gilbert-Elliot erasure channel," in *International Symposium on Information Theory*. Austin, TX: IEEE, June 2010.

[10] J.-F. Chamberland, H. D. Pfister, and S. Shakkottai, "First-passage time analysis for digital communication over erasure channels with delay-sensitive traffic," in *Allerton Conference on Communication, Control, and Computing*. Monticello, IL: IEEE, September 2010.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer, 1998.

[12] H. S. Wang and N. Moayeri, "Finite state Markov channel – A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, February 1995.

[13] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[14] L. Wilhelmsson and L. B. Milstein, "On the effect of imperfect interleaving for the Gilbert-Elliott channel," *IEEE Trans. Commun.*, vol. 47, no. 5, pp. 681–688, May 1999.

[15] P. Parag, J.-F. Chamberland, H. D. Pfister, and K. R. Narayanan, "Code rate, queueing behavior and the correlated erasure channel," in *IEEE Information Theory Workshop on Information Theory*, Cairo, Egypt, January 2009.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[17] P. Lancaster and M. Tismenetsky, *The theory of matrices: with applications*, 2nd ed. Academic Press, 1985.