First-Passage Time and Large-Deviation Analysis for Erasure Channels With Memory

Santhosh Kumar, *Student Member, IEEE*, Jean-François Chamberland, *Senior Member, IEEE*, and Henry D. Pfister, *Senior Member, IEEE*

***Abstract*—This paper considers the performance of digital communication systems transmitting messages over finite-state erasure channels with memory. Information bits are protected from channel erasures using error-correcting codes; successful receptions of codewords are acknowledged at the source through instantaneous feedback. The primary focus of this research is on delay-sensitive applications, codes with finite block lengths, and, necessarily, nonvanishing probabilities of decoding failure. The contribution of this paper is twofold. A methodology to compute the distribution of the time required to empty a buffer is introduced. Based on this distribution, the mean hitting time to an empty queue and delay-violation probabilities for specific thresholds can be computed explicitly. The proposed techniques apply to situations where the transmit buffer contains a predetermined number of information bits at the onset of the data transfer. Furthermore, as additional performance criteria, large deviation principles are obtained for the empirical mean service time and the average packet-transmission time associated with the communication process. This rigorous framework yields a pragmatic methodology to select code rate and block length for the communication unit as functions of the service requirements. Examples motivated by practical systems are provided to further illustrate the applicability of these techniques.**

***Index Terms*—Block codes, communication systems, data communication, Markov processes, queuing analysis.**

I. Introduction

CONTEMPORARY communication systems must be designed to accommodate the multiple applications that compose today's digital landscape. In particular, mobile devices must meet the heterogeneous needs of various data flows in terms of delay tolerance and bandwidth requirements. On the Internet backbone, congestion is often prevented by overprovisioning. The large throughput and low latency of parallel optical lines provide a pragmatic solution that offers adequate network performance. This approach, combined with localized content distribution networks and edge throttling, is key in supporting delay-sensitive traffic over the Internet core. Unfortunately, a similar strategy cannot be applied to connect untethered devices, as wireless physical resources are limited and costly. The narrow usable spectrum and the broadcast nature of wireless environments limit the effective bandwidth of wireless access networks and, hence, demand the efficient management of available resources.

In this paper, we develop a mathematical framework that enables the optimal allocation of link resources for wireless systems in the context of delay-sensitive communication. Distinguishing features of the proposed methodology include the joint treatment of finite-state channels with memory and queuing behavior at the transmitter. The focus is on the first-passage time to an empty queue, and the methodology implicitly provides a distribution for the time it would take an additional packet to reach the head of the queue. This view is not only important for resource allocation and performance evaluation, but also offers a foundation for choosing among possible routes and distinct interfaces. From an abstract perspective, we introduce a formulation where time dependence, which impacts decoding failures, is captured meticulously. In contrast to block-fading models, this formulation allows the seamless optimization of parameters such as code rate and block length. This is instrumental in better understanding how these parameters affect the overall performance of delay-sensitive wireless connections.

Several contributions on the interplay between decisions at the physical layer and overall performance at the link layer can be found in the literature [1]–[4]. Notable approaches include the outage capacity [5], [6], a probabilistic performance criterion based on the marginal distribution of channel blocks; the effective capacity [7], [8] which captures the decay rate in buffer occupancy at the transmitter; and finite block-length analyses of wireless connections [9], [10]. Physical resources can be optimized to reduce average delay by carefully selecting advantageous modulation schemes and coding strategies [11], [12]. Multiobjective problem formulations have also been explored. For instance, the optimal tradeoff between power and delay has received attention in the past [13]. The joint treatment of queuing and error-control coding has been examined by simultaneously considering the effective capacity of a link and the error exponent of a code family [14], [15]. Markov models have been successfully employed in the queuing analysis of communication links with automatic repeat request (ARQ) [16], [17]. Finally, powerful asymptotic techniques based on large deviations and heavy traffic limits have been developed to handle real-time traffic over unreliable links [18], [19].

This study differs from previous contributions in that it relates queuing behavior, error control coding, and channel evolution

Manuscript received June 13, 2012; revised December 23, 2012; accepted March 27, 2013. Date of publication May 02, 2013; date of current version August 14, 2013. This work was supported by the National Science Foundation under Grants 0747363 and 0830696. This paper was presented in part at the 2010 and 2011 Allerton Conferences on Communication, Control, and Computing.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: santhosh.kumar@tamu.edu; chmbrlnd@tamu.edu; hpfister@tamu.edu).

Communicated by D. Guo, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2260593

without resorting to asymptotically long coding delays or rough approximations. Decoding performance at the receiver captures channel correlation within a block, while the queuing aspect of the problem is key in understanding the impact of time dependences among successive decoding attempts. Together, they provide an accurate assessment of overall system performance and lead to novel guidelines about efficient designs.

Furthermore, by focusing on the first-passage time to an empty queue [20], we are able to bypass the search for representative arrival processes. Rather, resource management can be performed adaptively based on current system conditions. Having a distribution for the hitting time to an empty buffer enables the computation of several pertinent performance criteria such as the probability of violating a completion deadline, the mean first-passage time to an empty queue, and Chernoff bounds. The proposed methodology is closely related to generating functions [21] and it works well for reasonably small initial buffer sizes, which are typical of communication systems subject to stringent delay restrictions. On the other hand, under large buffers, this technique becomes somewhat cumbersome. In this latter case, 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. Indeed, the concentration of empirical measures can be used to gracefully adjust delay sensitivity to the needs of real-time data flows by selecting the deviation threshold, i.e., the argument of the rate function [22]. Once a threshold is set, system parameters can be optimized according to this objective function and the resulting performance can be predicted accurately.

Throughout, we assume the availability of reliable acknowledgments using periodic feedback. We also assume that the transmitter and receiver share a common randomness, which permits the utilization of random binary codes. The remainder of this paper is organized as follows. Section II presents the channel model and the random coding scheme. The queuing aspect of the problem is developed in Section III. A large deviations perspective on the mean transmission time and the average service rate is offered in Section IV. The findings are supplemented by a discussion of pertinent criteria for performance evaluation, along with numerical examples. Concluding remarks and possible avenues of future research are exposed in Section VII.

II. System Model

Channel memory is one physical aspect of wireless communication in which we are particularly interested. From a queuing perspective, it is well known that correlation over time can drastically alter the stationary distribution of a queuing system [23], [24]. In a similar manner, channel memory can have a strong impact on overall performance, as it induces time-dependence in the service process at the transmitter. This phenomenon is especially important for delay-sensitive applications that require the reliable, ordered delivery of data streams. One class of models that captures such dependence is the collection of finite-state channels with memory [25]–[27]. System models derived from this class of channels are typically mathematically tractable, and they offer a natural mechanism to account for correlation over time. Moreover, insights acquired by studying erasure channels can often be translated to error channels or, at least, provide partial intuition about promising solutions for the latter, more challenging scenarios.

This paper revolves around a communication paradigm where information bits flow from a source to a destination. The transmitter is assumed to possess a message of a certain length at the onset of the data transfer, and forward error correction is employed to shield content from potential symbol erasures. At the beginning of a transmission, the leading information bits stored at the source are grouped into a segment, and redundancy is added to this message using block encoding. The resulting codeword is then sent over a finite-state erasure channel with memory. Contingent upon the channel realization, the destination can either retrieve the data contained in the transmitted codeword or it declares a decoding failure. Successful transmissions are acknowledged and the corresponding bits are then discarded from the source buffer. Otherwise, the leading information bits remain in the queue. We emphasize that, in this framework, the original data sequence is guaranteed to be transferred unaltered. However, the completion time of the queue-emptying process is a random variable that depends on the coding/decoding strategy adopted and on the realization of the channel.

A. Channel Abstraction

As indicated previously, we capture channel randomness and its impact on the communication link using a finite-state Markov process. Several pertinent communication scenarios can be modeled in this manner [28]–[30]. At a particular time instant, we assume that the channel can be in one of k states taking value in $\mathcal{C} = \{1, 2, \ldots, k\}$. State transitions over time form a Markov chain. We denote the corresponding transition probability matrix by

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

Entry b_{ij} in matrix $\mathbf{B}$ represents the conditional probability that, starting from state i, the channel transitions to state j. As such, $\mathbf{B}$ is a right stochastic matrix. When in state i, the transmitted symbol is erased with probability ε_i and, consequently, it is received correctly with probability $1 - \varepsilon_i$. For notational convenience, we impose a quality ordering on the channel states, i.e., $\varepsilon_i \geq \varepsilon_j$ whenever $i < j$. We represent the state of the channel at time instant n by C_n. We note that $\{C_n\}$ is a first-order Markov process. A diagram illustrating the operation of the communication link for a two-state channel appears in Fig. 1.

Assumption 1: Throughout, we hypothesize that the chain governing the finite-state channel is irreducible and aperiodic. We also assume that this Markov channel is nontrivial in that there exists a state $i \in \mathcal{C}$ such that $\varepsilon_i < 1$.

As we shall see, these conditions guarantee the existence of a random coding scheme for which the transmission process terminates in finite time, almost surely. These transmission schemes are the only ones of interest for our purpose. In that

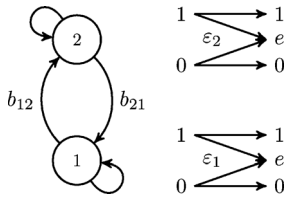

Fig. 1. Communication at the bit level takes place over a finite-state erasure channel with memory. While in state i, the probability of a bit erasure is ε_i. The evolution of the channel over time forms a Markov process.

sense, Assumption 1 is introduced to prevent difficulties that arise from idiosyncratic, irrelevant scenarios.

B. Coding Scheme

The envisioned system employs forward error correction to counteract possible channel erasures. A codeword transmission attempt is initiated by selecting the leading K bits from the source buffer. Redundancy is then added to this data segment through the encoding process. A random coding scheme is adopted as a mathematically convenient abstraction to realistic implementations [1], [31]. To create each codeword transmission, a random binary parity-check matrix of size $(N-K)\times N$ is generated. Every entry is selected uniformly over the binary alphabet, independently from other elements. The resulting codebook corresponds to the null space of this matrix. Such a coding scheme ensures that successful decoding of different codewords is conditionally independent given the channel states at the respective transmission times. This will greatly simplify the ensuing analysis. We assume that maximum-likelihood decoding is performed at the receiver.

We emphasize that this mode of operation requires shared randomness at the source and the destination. Interestingly, this coding scheme is known to perform well for large block lengths; and it supports flexible rates of communication, any rate of the form K/N where $0 \leq K \leq N$ is admissible. These random codes have the additional property that the average probability of decoding failure depends only on the number of erasures caused by the channel and not on the specific locations of these erasures. Provided that e erasures have occurred during transmission, the probability of decoding failure can be evaluated explicitly

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

A proof for this statement is based on the equivalence between the linear independence of the e erased columns in the parity-check matrix and the event of a successful decoding [31]. Throughout this paper, $P_{\mathrm{f}}(p,e)$ denotes

$$P_{\mathrm{f}}(p,e) = \begin{cases} 1-\prod_{l=0}^{e-1}\left(1-2^{l-p}\right) & \text{if } e \leq p \\ 1 & \text{if } p < e \leq N \end{cases} \tag{2}$$

which is the average probability of decoding failure under maximum likelihood of a codebook generated by using a random binary parity-check matrix of size $p\times N$, for any $N \geq p$, when e erasures have occurred.

C. Distribution of Erasures

From the earlier discussion, we gather that the number of erasures suffered by a codeword plays a critical role in determining overall system performance, as it dictates the probability of decoding failure. This random variable thus warrants due attention. Let E denote the number of erasures occurring in a given packet transmission. Since the probability of decoding failure of a codeword depends only on the number of erasures, it suffices to consider probabilities of the form $\Pr(E=e, C_{N+1}=j|C_1=i)$ to characterize the evolution of the system. Note that C_1 and C_{N+1} correspond to the channel state transitions across the first codeword transmission. We can describe this distribution in a compact form using matrix generating functions. Define matrix $\mathbf{B}_x$ by

$$\mathbf{B}_x = \begin{bmatrix} b_{11}(1-\varepsilon_1+\varepsilon_1 x) & \cdots & b_{1k}(1-\varepsilon_1+\varepsilon_1 x) \\ b_{21}(1-\varepsilon_2+\varepsilon_2 x) & \cdots & b_{2k}(1-\varepsilon_2+\varepsilon_2 x) \\ \vdots & \ddots & \vdots \\ b_{k1}(1-\varepsilon_k+\varepsilon_k x) & \cdots & b_{kk}(1-\varepsilon_k+\varepsilon_k x) \end{bmatrix}.$$

Throughout this paper, $[\![x^n]\!]$ denotes the linear operator that maps a polynomial in $\Re[x]$ to the coefficient of x^n. For $e \in \mathbb{N}_0$ and $i,j \in \mathcal{C}$, one can show that [21]

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

where, in this case, E denotes the number of erasures over an interval of length N. The probability that Markov process $\{C_n\}$ coincides with a specific sequence of states is equal to the probability of a certain path through the matching trellis. Moreover, at each point in time, the probability of observing an erasure only depends on the current state. Consequently, taking the Nth power of matrix $\mathbf{B}_x$ is an efficient way to compute the aggregate conditional probability of observing exactly e erasures, given an initial probability distribution and an end state. In other words, $\mathbf{B}_x^N$ offers a way to simultaneously sum all the relevant paths through the trellis. It is also possible to compute such probabilities through nested sums [32], but the ensuing equations rapidly become cumbersome for large values of N and Markov chains with sizable state spaces.

Given initial state i and for a fixed final state j, we can apply the total probability theorem to compute the probability of decoding failure

$$\sum_{e=0}^{N} P_{\mathrm{f}}(N-K,e)\Pr\left(E=e, C_{N+1}=j|C_1=i\right). \tag{4}$$

These conditional probabilities, along with the progression of the channel states, underlie the evolution of the queuing system.

Remark 1: As a side note, it is instructive to point out that, under Assumption 1, there exist values for N and K such that the probability of decoding success as a function of C_1 is not uniformly zero. In particular, if i is a channel state such that $\varepsilon_i < 1$, then for large enough N and $N-K$, the probability of decoding failure in (4) will be less than one. Random codes for which the conditional probability of decoding success is not uniformly zero are termed nontrivial.

III. Queuing Model

This section describes the queuing behavior of our system. First, we assume that the number of information bits present at the source at the beginning of the communication process is fixed and equal to ℓ. Given a code rate and block length, the source takes the leading K data bits and encodes the resulting segment into a codeword of length N using the scheme described in the preceding section. This codeword is then sent to the destination through N consecutive uses of the erasure channel. A service opportunity occurs every time the random code and channel realization jointly permit reliable decoding. We emphasize, again, that the destination is assumed to possess the ability to acknowledge the successful reception of codewords through instantaneous feedback. As such, the selected information bits remain in the transmit queue until a corresponding codeword is decoded faithfully at the destination. This data segment is immediately discarded from the buffer upon successful decoding of a packet.

In its simplest form, this scheme represents a variation of ARQ. We note that this mode of operation is somewhat naïve in that the information contained in failed decoding attempts is disregarded. A more astute implementation will seek to leverage past failures by performing joint decoding over all the observed messages pertaining to the current data segment. Incremental redundancy and hybrid ARQ are valuable techniques that can improve performance [33]–[35]. In this paper, we discuss both ARQ and its hybrid variant, where partial information from failed transmission attempts is incorporated in the decoding process. Still, we focus largely on the rudimentary scheme because it admits a simpler, more elegant characterization while preserving the natural tradeoff between error protection and payload content. Overall, the proposed methodology yields pertinent results that help improve our understanding of delay-sensitive systems.

Our primary interest lies in the distribution of the time elapsed until the message originally contained in the source buffer becomes wholly available at the destination. To capture this quantity adequately, we need to examine the evolution of the queue. The length of the queue can be expressed in terms of the number of data segments awaiting transmission. If a queue initially contains ℓ information bits, then it will require the successful reception of $m = \lceil \ell/K \rceil$ codewords until the last segment gets processed. The number of segments in the transmit buffer therefore becomes a measure of residual work until our objective is met, and it is intrinsically linked to the state of our communication system.

Codeword s denotes the block of transmitted bits during the time instants $sN + 1, \ldots, (s+1)N$, where $s \geq 0$. These codewords include both decoding successes and failures. For N fixed, we denote the size of the queue at the onset of codeword s by Q_s. We note that the state of the bit-erasure channel at the same time instant is C_{sN+1}. Thus, at the onset of the first codeword transmission ($s = 0$), the size of the queue is Q_0 and the state of the bit-erasure channel is C_1. The rapid succession of symbols in the bit-erasure channel compared to events taking place in the queue produces the mismatch in indexing between Q_s and C_{sN+1}. Indeed, queue transitions are only possible at the completions of decoding attempts, which only occur after every N symbol transmissions. The resulting stochastic process $\{Q_s\}$ is a hidden Markov process, as it is determined partly by the evolution of the unobserved channel process $\{C_n\}$. While $\{Q_s\}$ alone does not possess the Markov property, it is possible to create an augmented process containing Q_s with this desirable attribute. The particulars of the procedure depend on whether one is considering the standard ARQ framework or its hybrid variant. We treat these two instances separately.

A. ARQ

As the title suggests, this section focuses exclusively on the scenario where the source and the destination employ ARQ to overcome channel erasures and, thereby, achieve reliable data transmission. In particular, the information contained in past decoding attempts is disregarded by the decoder when receiving the latest codeword. To build a suitable model, we consider the random vector $U_s = (C_{sN+1}, Q_s)$ composed of channel state and queue length. We wish to show that this vector contains all the relevant information to track the evolution of the system.

Theorem 1: The aggregate process $\{U_s\}_{s\geq 0}$ possesses the Markov property. That is, conditioned on $U_t = (i, q)$, the stochastic process $\{U_{s+t}\}_{s\geq 0}$ is independent of $U_0, \ldots, U_{t-1}$.

Proof: See Appendix A. ∎

Using the total probability theorem, we can write the transition probabilities of $\{U_s\}$ as follows:

$$\begin{aligned}&\Pr\left(U_{s+1} = (j, q_{s+1}) | U_s = (i, q_s)\right)\\ &\quad = \sum_{e=0}^{N} \Pr\left(Q_{s+1} = q_{s+1} | E = e, Q_s = q_s\right) \times\\ &\qquad \Pr\left(E = e, C_{(s+1)N+1} = j | C_{sN+1} = i\right) \end{aligned} \tag{5}$$

where $i, j \in \mathcal{C}$. For a nonempty queue, the first part of each summand corresponds to one of three possible cases

$$\begin{aligned}&\Pr\left(Q_{s+1} = q_{s+1} | E = e, Q_s = q_s\right)\\ &\quad = \begin{cases} P_{\mathrm{f}}(N-K, e), & q_{s+1} = q_s\\ 1 - P_{\mathrm{f}}(N-K, e), & q_{s+1} = q_s - 1\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

The probability of decoding failure $P_{\mathrm{f}}(\cdot, \cdot)$ appears in (1), while the conditional distribution of erasures within a block is given in (3). Thus, we have already developed the tools necessary to efficiently compute the value of every transition probability in (5). The evolution of the queuing system and its admissible transitions are depicted graphically in Fig. 2. The states $\{(\cdot, q)\}$ are collectively referred to as the qth level of the queue. The first-passage time to an empty buffer is therefore equivalent to the hitting time to level zero. Due to the repetitive structure of this augmented system, the hitting time to a lower level will play a key role in finding a tractable solution to the problem at hand.

An additional quantity of interest in the analysis of delay-sensitive systems is the mean service rate. To compute this quantity, it is convenient to analyze the service process $\{D_s\}$, where D_s indicates the potential of a successful decoding event at time s, $s \geq 0$. That is, $D_s = 1$ when a message is (or could

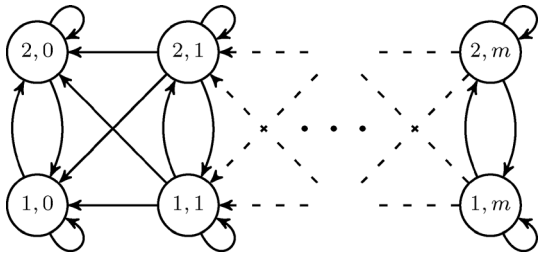

Fig. 2. This figure illustrates the progression of the queuing system for a service process that is governed by a two-state Markov erasure channel. System states, which are composed of queue lengths and channel states, are represented by circles. Admissible transitions are marked by the arrows.

be) decoded faithfully at the destination; and $D_s = 0$ otherwise. In words, the sequence $\{D_s\}$ indicates time instants at which blocks of information are (or could be) transferred successfully to the destination. As in the case of the queuing abstraction, the stochastic process $\{D_s\}$ forms a hidden Markov process which can be lifted to an augmented Markov process. Let $V_s = \left(C_{(s+1)N+1}, D_s\right)$ denote a random vector 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. As in Theorem 1, one can show that the stochastic process $\{V_s\}$ forms a Markov chain.

We note that the transition probabilities of $\{D_s\}$ are closely related to those of $\{Q_s\}$. Since there are no arrivals in our framework, the evolution of these processes are governed by

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

For convenience, we establish a succinct notation for the transition probabilities of our two augmented processes,

$$\begin{aligned}\kappa_{ij} &= \Pr(U_{s+1} = (j,q) | U_s = (i,q)) \\ &= \Pr(V_{s+1} = (j,0) | V_s = (i,d)) \\ \mu_{ij} &= \Pr(U_{s+1} = (j,q-1) | U_s = (i,q)) \\ &= \Pr(V_{s+1} = (j,1) | V_s = (i,d)) \end{aligned} \tag{6}$$

where $q \in \mathbb{N}$, $i, j \in \mathcal{C}$ and $d \in \{0,1\}$. These common definitions draw further attention to the close connection between $\{U_s\}$ and $\{V_s\}$.

In view of Remark 1 and for nontrivial codes, there exists $i \in \mathcal{C}$ such that $\mu_{ij} > 0$. This implies that the states associated with an empty buffer form the only closed communicating class and, as such, the remaining states are transient [20]. Since the number of states in the augmented chain is finite, this structure ensures that the task of emptying the transmit buffer is carried out in finite time, almost surely.

The symmetric decomposition of the queuing system into levels suggests an approach based on the quasi-birth-death structure of the chain. Suppose that the buffer contains exactly m data segments at time zero, i.e., $Q_0 = m$. We can define the hitting time from level m to level q of the chain as

$$H_q = \inf\{s \geq 0 | Q_s = q\}, \tag{7}$$

where $0 \leq q < m$. That is, H_q designates the time instant at which the process $\{U_s\}$ first enters the qth level of the queue. We emphasize that, under the mild assumptions discussed above, H_q is almost surely finite. For consistency, we also define $H_m = 0$. Noting that Q_s is a nonincreasing process, we can write the sojourn time at level q as

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

where $0 < q \leq m$. That is, random variable T_q denotes the amount of time $\{U_s\}$ stays at level q before leaving for the subsequent lower level.

We are especially interested in H_0, the first-passage time to an empty queue. Taking advantage of the structure of the augmented Markov chain, we can fragment H_0 into a sum of elementary components. Specifically, the hitting time H_0 is equal to the sum of the sojourn times $T_1, \ldots, T_m$, i.e.,

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

The sojourn times T_q and T_{q-1} are coupled through the channel state $C_{NH_{q-1}+1}$ and hence are not independent. However, since the codebooks over different codeword transmissions are independent, the sojourn times $T_1, \ldots, T_m$ are conditionally independent given the channel states $\left\{C_{NH_q+1}\right\}_{q=0}^{m}$. The sojourn times $T_1, \ldots, T_m$ are also conditionally identically distributed. That is

$$\Pr\left(T_q = t, C_{NH_{q-1}+1} = j | C_{NH_q+1} = i\right)$$

is independent of q. A powerful means to compute the distribution of H_0 is to employ generating functions extended to matrices [21], exploiting the conditional independence and the identical distribution among the sojourn times $\{T_q\}$. This more intricate version of the generating function is necessary to keep track of the channel state entered after each downward queue transition. This method is described below.

Consider a reduced Markov chain composed of states $\{(i,0),(i,1)\}_{i=1}^{k}$, as shown in Fig. 3 for a Gilbert–Elliott channel. This reduced Markov chain represents one downward queue transition of the original system. Under proper state ordering, we can write the transition probability matrix for the reduced subsystem as

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{M} & \mathbf{K} \end{bmatrix}, \tag{8}$$

where we have implicitly defined matrices

$$\mathbf{M} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1k} \\ \mu_{21} & \cdots & \mu_{2k} \\ \vdots & \ddots & \vdots \\ \mu_{k1} & \cdots & \mu_{kk} \end{bmatrix} \quad \mathbf{K} = \begin{bmatrix} \kappa_{11} & \cdots & \kappa_{1k} \\ \kappa_{21} & \cdots & \kappa_{2k} \\ \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} \end{bmatrix}.$$

We emphasize that $\mathbf{P}$ is a stochastic matrix. As a consequence of the Perron–Frobenius theorem, we know that the spectral radius associated with $\mathbf{P}$ is one [36].

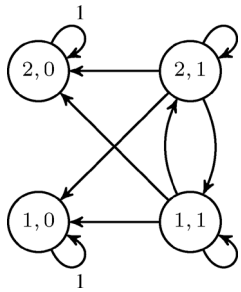

Fig. 3. This reduced Markov diagram represents one of the quasi-birth-death subcomponents of the queuing system. Starting from any distribution over these four states, it is possible to characterize the sojourn time T spent at level one. This is a key step in deriving the first-passage time to an empty buffer.

Define sojourn time T as the time spent at queue-level 1 of the reduced Markov chain. Mimicking our original notation, let Q_s denote the level of the queue (either 1 or 0) at the onset of codeword s and let $U_s = (C_{sN+1}, Q_s)$. Suppose the reduced Markov chain starts at queue-level 1, i.e., $Q_0 = 1$, then

$$T = \inf\{s \geq 0 | Q_s = 0\}.$$

The random variables $\{T_q\}_{q=1}^m$ and T have identical conditional distributions. That is, for any $1 \leq q \leq m$

$$\begin{aligned}&\Pr(T = t, C_{NT+1} = j | C_1 = i)\\ &\quad = \Pr\left(T_q = t, C_{NH_{q-1}+1} = j | C_{NH_q+1} = i\right).\end{aligned}$$

The distributions of the sojourn times $T_1, \ldots, T_m$ are important for determining the distribution of H_0. Thus, the above relation between $T_1, \ldots, T_m$ and T implies that the distribution of T is critical. Generating functions are an elegant way to characterize such distributions. Define matrix generating function $\mathbf{G}_T(z)$ entrywise by

$$[\mathbf{G}_T(z)]_{ij} = \mathrm{E}\left[z^T \mathbf{1}_{\{C_{NT+1}=j\}} | C_1 = i\right] \tag{9}$$

where $\mathbf{1}_{\{\cdot\}}$ is the standard set indicator function.

Lemma 1: For the reduced subsystem associated with (8), the matrix generating function $\mathbf{G}_T(z)$ is equal to

$$\mathbf{G}_T(z) = (\mathbf{I} - \mathbf{K}z)^{-1}\mathbf{M}z. \tag{10}$$

Proof: The matrix generating function $\mathbf{G}_T(z)$ can be obtained by treating the entries of $\mathbf{P}$ as real polynomials in z, with

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

Consider the two states $(i,1)$ and (j,l), where $l = 0$ or $l = 1$. Their indices in the ordering associated with $\mathbf{P}$ are $k + i$ and $lk + j$, respectively. Recall that $[\![z^t]\!]$ denotes the operator that maps a polynomial in z to the coefficient of z^t. Suppose that, at time zero, the reduced system starts in state $(i,1)$. After s transmissions, the reduced system will be in state $(j,1)$ only when all the s transmissions result in decoding failures. Thus

$$\Pr(U_s = (j,1) | U_0 = (i,1)) = [\mathbf{K}^s]_{i,j}. \tag{11}$$

Similarly, the probability that the reduced system is in state $(j,0)$ after s transmissions and having spent exactly t steps in queue-level 1, where $1 \leq t \leq s$, is given by

$$\sum_{h=1}^{k} \Pr(U_s = (j,0), U_t = (j,0), U_{t-1} = (h,1) | U_0 = (i,1)).$$

Since the reduced system does not transition to a different state after reaching queue-level 0 (see Fig. 3), this can be reduced to

$$\begin{aligned}&\sum_{h=1}^{k} \Pr(U_s = (j,0), U_t = (j,0), U_{t-1} = (h,1) | U_0 = (i,1))\\ &\quad = \sum_{h=1}^{k} \Pr(U_t = (j,0), U_{t-1} = (h,1) | U_0 = (i,1))\\ &\quad = \left[\mathbf{K}^{t-1}\mathbf{M}\right]_{i,j}.\end{aligned} \tag{12}$$

Combining (11) and (12), the joint probability that the reduced system is in state (j,l) at time $s > 0$ and has spent exactly t steps at queue-level 1, where $1 \leq t \leq s$, can be expressed compactly as

$$\Pr(S_s = t, U_s = (j,l) | U_0 = (i,1)) = \left[\!\left[z^t\right]\!\right] [\mathbf{P}_z^s]_{k+i,lk+j},$$

where S_s represents the total time spent at queue-level 1 over the interval from zero to instant s. Since T is a discrete random variable that is finite almost surely

$$\begin{aligned}[\mathbf{G}_T(z)]_{ij} &= \mathrm{E}\left[z^T \mathbf{1}_{\{C_{NT+1}=j\}} | C_1 = i\right]\\ &= \lim_{s\to\infty} \sum_{t=0}^{s} \Pr(T = t, C_{Nt+1} = j | C_1 = i)\, z^t\\ &= \lim_{s\to\infty} \sum_{t=0}^{s} \Pr(S_s = t, U_s = (j,0) | U_0 = (i,1))\, z^t\\ &= \lim_{s\to\infty} \sum_{t=0}^{s} \left(\left[\!\left[z^t\right]\!\right] [\mathbf{P}_z^s]_{k+i,j}\right) z^t\\ &= \lim_{s\to\infty} [\mathbf{P}_z^s]_{k+i,j}.\end{aligned}$$

Therefore, the generating matrix $\mathbf{G}_T(z)$ can be obtained as

$$\begin{aligned}\mathbf{G}_T(z) &= \lim_{s\to\infty} \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \mathbf{P}_z^s \begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\\ &= \lim_{s\to\infty} \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \sum_{t=1}^{s} \mathbf{K}^{t-1}\mathbf{M}z^t & \mathbf{M}^s z^s\end{bmatrix} \begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\\ &= \lim_{s\to\infty} \sum_{t=1}^{s} \mathbf{K}^{t-1}\mathbf{M}z^t\\ &= (\mathbf{I} - \mathbf{K}z)^{-1}\mathbf{M}z.\end{aligned}$$

The above equation holds for all $|z| < \varrho(\mathbf{K})^{-1}$, where $\varrho(\cdot)$ denotes the spectral radius of its matrix argument. ■

B. Hybrid ARQ

Hybrid ARQ is a mechanism that seeks to incorporate the partial information contained in failed transmissions into the subsequent decoding attempts of the same data segment. In this sense, it differs significantly from ARQ only when the initial

decoding of a data segment fails. For finite-state erasure channels with memory, the evolution of a hybrid ARQ system can be characterized completely, although in a somewhat cumbersome manner. To implement hybrid ARQ with random codes, we must modify our coding strategy slightly.

Herein, we focus on hybrid schemes with finite depths. That is, the transmitter–receiver pair has a predetermined number of tries to successfully transmit a data segment. Our favored implementation relies on puncturing random codes. In a way analogous to our previous approach, we generate a codebook by creating a random binary parity-check matrix of size $(aN-K)\times aN$, where a is the depth of the hybrid ARQ scheme. Again, the entries are selected uniformly from the binary alphabet and the codebook is equal to the null space of this matrix. The hybrid ARQ scheme progresses as follows. First, an information segment is mapped to a codeword of length aN. During the initial transmission, the leading N symbols of this codeword are sent over the erasure channel. Upon completion of this phase, the destination tries to recover the original data segment. When decoding fails, the next N symbols are sent and the aggregate message is run through a maximum-likelihood decoder. This process continues, communicating N symbols at a time, until the message is successfully decoded at the destination or the total number of attempts reaches its limit.

Since untransmitted symbols can be classified as erasures for the purpose of decoding, we can leverage (2) in assessing the probabilities of decoding failure at the destination. That is, when s codeword chunks are present at the destination, out of which a total of e symbols are erased, the probability of decoding failure can be written as

$$P_{\mathrm{f}}(aN-K,e+(a-s)N) = 1-\prod_{i=0}^{e+(a-s)N-1}\left(1-2^{i-(aN-K)}\right). \tag{13}$$

Comparing this expression for $s=1$ and $a>1$ to (1), we gather that the probability of decoding failure after receiving one chunk of length N for the hybrid ARQ scheme differs from the probability of failure in standard ARQ. Indeed, there is a slight penalty for the initial transmission resulting from using a random code tailored to hybrid ARQ. The following proposition establishes a uniform bound on the loss in performance associated with the hybrid scheme.

Proposition 1: Suppose that p and e are fixed, positive integers. The function of n defined by

$$P_{\mathrm{f}}(p+n,e+n)=\begin{cases}1-\prod_{l=0}^{n+e-1}\left(1-2^{l-p-n}\right) & \text{if } e\le p\\ 1 & \text{if } e>p\end{cases}$$

is monotone increasing. Furthermore, the difference between this function and $P_{\mathrm{f}}(p,e)$ is uniformly bounded,

$$P_{\mathrm{f}}(p+n,e+n)-P_{\mathrm{f}}(p,e)\le 2^{-p}.$$

Proof: See Appendix B. ■

The probability of decoding failure for the initial transmission of the hybrid ARQ scheme is $P_{\mathrm{f}}(aN-K,e+(a-1)N)$, and it is $P_{\mathrm{f}}(N-K,e)$ for the standard ARQ scheme when the codeword suffers e erasures. As an immediate consequence of Proposition 1, we know that the penalty incurred in using hybrid ARQ in terms of probability of decoding failure at the first attempt is

$$P_{\mathrm{f}}(aN-K,e+(a-1)N)-P_{\mathrm{f}}(N-K,e)\le 2^{-(N-K)},$$

which remains very small for typical scenarios. This brings credibility to employing a punctured random code in our analysis.

Using random codes over erasure channels leads to some highly desirable properties for the hybrid ARQ problem. These properties are, in turn, instrumental in finding expressions for the probabilities of success at intermediate decoding attempts. Suppose that a codebook is generated using a $(aN-K)\times aN$ parity-check matrix. For this specific code, if decoding fails given the first sN received symbols (including erasures), then it will necessarily be impossible to decode the message using the leading $(s-1)N$ received symbols. This nesting is in stark contrast to error channels.

We employ $P_{\mathrm{f}}^{(s)}(j|i)$ and $P_s^{(s)}(j|i)$ to denote the conditional probability of decoding failure and first reliable decoding success at attempt s, respectively, with final state j and given initial state i. The conditional probabilities of decoding failure are equal to

$$P_{\mathrm{f}}^{(s)}(j|i)=\sum_{e=0}^{sN}P_{\mathrm{f}}(aN-K,e+(a-s)N)\times \Pr\left(E_{sN}=e,C_{sN+1}=j|C_1=i\right).$$

Above, E_{sN} represents the number of erasures over the discrete interval $[1,sN]$. Given the probabilities of failure events, the conditional probabilities of success can be evaluated in a recursive fashion. Since decoding failure and decoding success at attempt one are complementary events, we have

$$\Pr(C_{N+1}=j|C_1=i)=P_{\mathrm{f}}^{(1)}(j|i)+P_s^{(1)}(j|i).$$

Thus, the probability of a success at time one with final state j given initial state i can be written as

$$P_s^{(1)}(j|i)=\Pr(C_{N+1}=j|C_1=i)-P_{\mathrm{f}}^{(1)}(j|i).$$

We note that this equation is the complement of (4), with a convenient new notation and appropriate parameters.

Similarly, consider the first two attempts in a hybrid ARQ scheme. Three disjoint events can occur: decoding failure at attempt two, decoding success for the first time at attempt two, decoding success at attempt one after which the channel enters some state l. Summing over all intermediate states l

$$\Pr(C_{2N+1}=j|C_1=i)=P_{\mathrm{f}}^{(2)}(j|i)+P_s^{(2)}(j|i)+\sum_{l\in\mathcal{C}}P_s^{(1)}(l|i)\Pr(C_{2N+1}=j|C_{N+1}=l).$$

Consequently, the conditional probability of being able to decode for the first time at attempt two with final state j and under initial state i is

$$P_s^{(2)}(j|i)=\Pr(C_{2N+1}=j|C_1=i)-P_{\mathrm{f}}^{(2)}(j|i)-\sum_{l\in\mathcal{C}}P_s^{(1)}(l|i)\Pr(C_{2N+1}=j|C_{N+1}=l).$$

Extending this procedure, the probability of a decoding success at attempt s and final state j, conditioned on initial state i, is given by

$$P_s^{(s)}(j|i) = \Pr\left(C_{sN+1} = j | C_1 = i\right) - P_{\mathrm{f}}^{(s)}(j|i) - \sum_{r=1}^{s-1} \sum_{l \in \mathcal{C}} P_s^{(r)}(l|i) \Pr\left(C_{sN+1} = j | C_{rN+1} = l\right).$$

This methodology provides a recursive and efficient way to compute the probabilities that, under hybrid ARQ, a system takes exactly s coded chunks to decode the original message. As in Section III-A, we intend to compute the matrix generating function of T, the time spent in the first level of the reduced Markov chain.

Consider the aforementioned hybrid ARQ scheme with depth equal to a. When there is a decoding failure at attempt a, the hybrid ARQ system has a few potential options. The system can discard previously received symbols altogether and start the process anew. Alternatively, the transmitter can re-encode the data segment and the information in previously received symbols can be used as side information during the decoding process. No matter what the exact strategy is, the queue occupancy of a hybrid ARQ system can always be lower and upper bounded.

1) *Lower bound*: In this mode, the decoding of a message always succeeds by the ath attempt. We call this the optimistic system. Let $\check{T}$ denote the time spent in the first level of the reduced Markov chain associated with this system.
2) *Upper bound*: In this mode, whenever decoding fails at the ath attempt, previously received symbols are discarded altogether and the process starts anew. We call this the pessimistic view. Let $\hat{T}$ denote the time spent in the first level of the reduced Markov chain of this system.

In essence, $\check{T}$ and $\hat{T}$ are Markov times that provide lower and upper bounds on T, the true stopping time of the hybrid ARQ decoding process. These strategies jointly produce a near complete characterization of the behavior of hybrid ARQ systems. We turn to the specifics of the proposed approaches below.

As mentioned previously, an optimistic bound (lower bound) on T can be derived using

$$\hat{P}_s^{(a)}(j|i) = \Pr\left(C_{aN+1} = j | C_1 = i\right) - \sum_{r=1}^{a-1} \sum_{l \in \mathcal{C}} P_s^{(r)}(l|i) \Pr\left(C_{aN+1} = j | C_{rN+1} = l\right),$$

instead of $P_s^{(a)}(j|i)$, by assuming that the decoding always succeeds by the ath attempt. This bound holds irrespective of how the system handles failures at attempt a. We define the optimistic matrix generating function $\mathbf{G}_{\check{T}}(z) = \mathbf{G}_{\min\{T,a\}}(z)$ entrywise by

$$\left[\mathbf{G}_{\check{T}}(z)\right]_{ij} = \sum_{r=1}^{a-1} P_s^{(r)}(j|i) z^r + \hat{P}_s^{(a)}(j|i) z^a.$$

The pessimistic matrix generating function $\mathbf{G}_{\hat{T}}(z)$ can be derived in two steps. First, consider the matrix generating function

$$\left[\mathbf{G}_{\tilde{T}}(z)\right]_{ij} = \sum_{r=1}^{a} P_s^{(r)}(j|i) z^r.$$

Then, under the assumption that information is discarded when the a decoding attempts have failed, we get

$$\mathbf{G}_{\hat{T}}(z) = \sum_{t=0}^{\infty} z^{at} \left(\mathbf{P}_{\mathrm{f}}^{(a)}\right)^t \mathbf{G}_{\tilde{T}}(z) = \left(\mathbf{I} - z^a \mathbf{P}_{\mathrm{f}}^{(a)}\right)^{-1} \mathbf{G}_{\tilde{T}}(z).$$

Above, the matrix $\mathbf{P}_{\mathrm{f}}^{(a)}$ is defined entrywise as

$$\left[\mathbf{P}_{\mathrm{f}}^{(a)}\right]_{ij} = P_{\mathrm{f}}^{(a)}(j|i).$$

We will return to these bounds and their application in Section VI.

C. Hitting Time to an Empty Buffer

We can build upon the matrix generating function of T to obtain the distribution of H_0. The basic insights behind this characterization are that the sojourn time at any level is finite almost surely and generating matrices can account for conditional independence.

Theorem 2: The ordinary generating function of H_0, the first-passage time to an empty queue, is given by

$$G_{H_0}(z) = \mathrm{E}\left[z^{H_q}\right] = \pi_0 \left(\mathbf{G}_T(z)\right)^m \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{14}$$

where π_0 is the channel state probability vector at time zero.

Proof: This expression for $G_{H_0}(z)$ can be obtained from an application of mathematical induction, which proceeds backward in time. The first step consists in showing that the hypothesis holds for the base case, the sojourn time at level m,

$$\begin{aligned} \left[\pi_0 \mathbf{G}_{T_m}(z)\right]_j &= \sum_{i=1}^{k} \left[\mathbf{G}_{T_m}(z)\right]_{ij} \Pr(C_1 = i) \\ &= \sum_{i=1}^{k} \mathrm{E}\left[z^{T_m} \mathbf{1}_{\{C_{NT_m+1}=j\}} \middle| C_1 = i\right] \Pr(C_1 = i) \\ &= \mathrm{E}\left[z^{T_m} \mathbf{1}_{\{C_{NT_m+1}=j\}}\right] = \mathrm{E}\left[z^{H_{m-1}} \mathbf{1}_{\left\{C_{NH_{m-1}+1}=j\right\}}\right] \end{aligned}$$

where we have used the fact that $H_{m-1} = T_m$. Thus, we gather that

$$\left[\left[z^t\right]\right] \left[\pi_0 \mathbf{G}_{T_m}(z)\right]_j = \Pr\left(H_{m-1} = t, C_{NH_{m-1}+1} = j\right).$$

We continue with the inductive step in a similar manner. Suppose that the hypothesis is true for a certain integer q where $0 < q < m$; that is,

$$\begin{aligned} \left[\pi_0 \mathbf{G}_{H_q}(z)\right]_j &= \mathrm{E}\left[z^{H_q} \mathbf{1}_{\{C_{NH_q+1}=j\}}\right] \\ &= \left[\pi_0 \mathbf{G}_{T_m}(z) \cdots \mathbf{G}_{T_{q+1}}(z)\right]_j. \end{aligned}$$

Then, we can write

$$\begin{aligned}
&\mathrm{E}\left[z^{H_{q-1}}\mathbf{1}_{\{C_{NH_{q-1}+1}=j\}}\right] = \mathrm{E}\left[z^{H_q+T_q}\mathbf{1}_{\{C_{NH_{q-1}+1}=j\}}\right] \\
&= \sum_{i=1}^{k} \mathrm{E}\left[z^{H_q+T_q}\mathbf{1}_{\{C_{NH_{q-1}+1}=j\}}\,\middle|\, C_{NH_q+1}=i\right] \times \\
&\quad \Pr(C_{NH_q+1}=i) \\
&= \sum_{i=1}^{k} \mathrm{E}\left[z^{H_q}\mathbf{1}_{\{C_{NH_q+1}=i\}}\right] \times \\
&\quad \mathrm{E}\left[z^{T_q}\mathbf{1}_{\{C_{NH_{q-1}+1}=j\}}\,\middle|\, C_{NH_q+1}=i\right] \\
&= \sum_{i=1}^{k} \left[\pi_0\mathbf{G}_{T_m}(z)\cdots\mathbf{G}_{T_{q+1}}(z)\right]_i \left[\mathbf{G}_{T_q}(z)\right]_{ij} \\
&= \left[\pi_0\mathbf{G}_{T_m}(z)\cdots\mathbf{G}_{T_q}(z)\right]_j = \left[\pi_0\mathbf{G}_{H_{q-1}}(z)\right]_j .
\end{aligned}$$

That is, the hypothesis is also true for $q-1$. We note that the third equality follows from the conditional independence of our quasi-birth-death Markov process. In our problem, we have $\mathbf{G}_{T_q}(z) = \mathbf{G}_T(z)$ for all $q \in \{1,\ldots,m\}$. Since this expression holds for any π_0, we conclude that $\mathbf{G}_{H_0}(z) = (\mathbf{G}_T(z))^m$ and as a consequence

$$\left[\left[z^t\right]\right]\left[\pi_0\left(\mathbf{G}_T(z)\right)^m\right]_j = \Pr\left(H_0 = t, C_{NH_0+1} = j\right).$$

Summing over all the possible end states, we recover the expression for $G_{H_0}(z)$ given in (14). ■

To differentiate among possible initial conditions, it will become useful to write the first-passage time to an empty queue with an initial buffer size of m segments as $H_0^{(m)}$.

IV. Large Deviation Analysis

As seen in the previous section, it is possible to evaluate the exact distribution of $H_0^{(m)}$. This facilitates the selection of parameters to optimize overall performance. However, this process becomes cumbersome for large buffer sizes. In such circumstances, analyzing the large deviations governing the system offers a new direction to derive meaningful guidelines for resource allocation and parameter tuning. Below, we study two types of aberrations under the ARQ scheme: deviations in the average transmission time and the mean service rate. We note that, although large deviations can be studied under hybrid ARQ, this latter scenario is somewhat tedious and it offers limited additional insights. Hence, we restrict our attention to the ARQ scheme. We begin with the average transmission time, that is, the normalized first-passage time to an empty queue.

A. Normalized First-Passage Time

Again, suppose that the transmit buffer contains exactly m segments at the onset of the communication process. 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 for Y_m is

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

The existence of limits of properly scaled logarithmic moment generating functions suggests that $\{Y_m\}$ may satisfy a large deviation principle [22]. In particular, consider the following asymptotic regime:

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

A few observations concerning $\Lambda(\lambda)$ are in order. In view of Lemma 1 and for $z = e^{\lambda}$

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

is a nonnegative matrix over the extended real numbers. In fact, this matrix possesses additional properties that are summarized below. Again, let $\varrho(\cdot)$ denote the spectral radius of its matrix argument.

Lemma 2: If T is finite almost surely, the matrix generator $\mathbf{G}_T\left(e^{\lambda}\right)$ exists as a nonnegative real matrix if and only if $\lambda < -\log\varrho(\mathbf{K})$. In particular, when $\lambda \geq -\log\varrho(\mathbf{K})$, one or more entries of $\mathbf{G}_T(e^{\lambda})$ will be infinite.

Proof: See Appendix C. ■

Another important quantity is the spectral radius of $\mathbf{K}$, which is related to the support of $\mathbf{G}_T\left(e^{\lambda}\right)$ as seen in Lemma 2.

Corollary 1: If T is finite almost surely, then $\varrho(\mathbf{K}) < 1$.

Proof: See Appendix D. ■

Under Assumption 1 and for any nontrivial coding scheme, T is finite almost surely; thus, the hypotheses of Lemma 2 and Corollary 1 are satisfied. A sufficient condition to ensure the existence of a large deviation principle for the average transmission time is that the Markov process $\{U_t\}$ sampled at departure events $\{H_q\}$ is irreducible. This guarantees that the states of the corresponding jump chain form a unique recurrent class. Formally, we postulate the following condition.

Assumption 2: The matrix $(\mathbf{I}-\mathbf{K})^{-1}\mathbf{M}$ is irreducible.

We note that, strictly speaking, this is not a necessary condition. Having a unique communicating class and, possibly, transient states in the jump chain will also work. However, this more encompassing setting leads to extra bookkeeping, which unnecessarily clouds some of the underlying concepts. Furthermore, all the practical systems we wish to study fulfill the requirements of Assumption 2. As such, we take it for granted from this point forward. Under this assumption, the matrix $\mathbf{G}_T\left(e^{\lambda}\right)$ is irreducible for any $\lambda < -\log\varrho(\mathbf{K})$ and, hence, the Perron–Frobenius theorem applies [22], [36]. This leads to the following result.

Proposition 2: Under Assumption 2, the limiting moment generating function defined in (15) exists as an extended real number for every $\lambda \in \mathbb{R}$, with

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

Proof: See Appendix E. ∎

Using matrix norms, it can be shown that $\mathbf{G}_T\left(e^{\lambda}\right)$ is differentiable entrywise 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 λ (see, e.g., [37, Th. 11.5.1],[22, p. 75]). Corollary 1 asserts that $\varrho(\mathbf{K}) < 1$, which implies that $\Lambda(0)$ is finite. In view of the discussion above, we conclude that the origin is in the interior of $\{\lambda \in \mathbb{R} : \Lambda(\lambda) < \infty\}$. Consequently, $\Lambda(\lambda)$ is essentially smooth and the Gärtner–Ellis theorem applies [22], thereby establishing the desired result.

Theorem 3: Suppose $\left\{Y_m = \frac{1}{m}\sum_{q=1}^{m} T_q\right\}$ is 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 $c \in \mathcal{C}$

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

where Γ^o and $\bar{\Gamma}$ denote the interior and closure of the set Γ, respectively.

Example 1: For the Gilbert–Elliott channel shown in Fig. 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 in γ 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 $\mathbf{G}_T\left(e^{\lambda}\right)$. As such, $\Lambda^*(\cdot)$ can be obtained efficiently.

B. Empirical Mean Service

We turn to the second type of aberrations we wish to study: deviations in the empirical mean service rate

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

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 (i, d). Using this ordering, the transition probability matrix $\mathbf{\Pi}$ for the augmented process $\{V_s\}$ is given 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.

Assumption 3: The matrix $\mathbf{\Pi}$ is irreducible.

This assumption is similar in spirit to Assumption 2. Yet, the large deviation principle on the empirical service can be derived under weaker conditions. In particular, it suffices to show that $\mathbf{K}+\mathbf{M}$ is irreducible, a requirement that is easily met. We stress that $\mathbf{K} + \mathbf{M}$ is equal to $\mathbf{B}^N$, and the latter matrix is itself irreducible by Assumption 1.

Theorem 4 ([22]): Let $\{V_s\}$ be a finite-state Markov chain possessing an irreducible transition matrix $\mathbf{\Pi}$. For every $x \in \mathbb{R}$, define

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \log \varrho\left(\mathbf{\Pi}_\lambda\right)\} \tag{17}$$

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

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

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^{π} 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 (6). We note that

$$\begin{aligned} &\Pr\left(V_{s+1} = (j, d_2) | V_s = (i, d_1)\right) \\ &\quad = \Pr\left(V_{s+1} = (j, d_2) | C_{(s+1)N+1} = i\right) \end{aligned}$$

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 written explicitly as

$$\mathbf{\Pi}_\lambda = \begin{bmatrix} \kappa_{11} & \cdots & \kappa_{1k} & \mu_{11}e^{\lambda} & \cdots & \mu_{1k}e^{\lambda} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} & \mu_{k1}e^{\lambda} & \cdots & \mu_{kk}e^{\lambda} \\ \kappa_{11} & \cdots & \kappa_{1k} & \mu_{11}e^{\lambda} & \cdots & \mu_{1k}e^{\lambda} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \kappa_{k1} & \cdots & \kappa_{kk} & \mu_{k1}e^{\lambda} & \cdots & \mu_{kk}e^{\lambda} \end{bmatrix}. \tag{18}$$

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

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

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

$$\begin{aligned} \det(\gamma\mathbf{I} - \mathbf{\Pi}_\lambda) &= \det\left((\gamma\mathbf{I} - \mathbf{M}e^\lambda)(\gamma\mathbf{I} - \mathbf{K}) - \mathbf{M}\mathbf{K}e^\lambda\right) \\ &= \det\left((\gamma\mathbf{I} - \mathbf{K})(\gamma\mathbf{I} - \mathbf{M}e^\lambda) - \mathbf{K}\mathbf{M}e^\lambda\right) \\ &= \det\left(\gamma^2\mathbf{I} - \gamma\mathbf{K} - \gamma\mathbf{M}e^\lambda\right). \end{aligned}$$

Collectively, Theorem 4 and the matrix defined in (18) 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 an example based on a two-state channel with memory.

Example 2: Once again, consider a Gilbert–Elliott erasure channel with $\mathcal{C} = \{1, 2\}$. The dimension of the state space in this case is $|\mathcal{V}| = 4$. Using the commutative block structure discussed above, the determinant of $(\gamma\mathbf{I} - \mathbf{\Pi}_\lambda)$ reduces to

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

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

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

For fixed parameters, this dominating root can be computed using the well-known quadratic formula. We will revisit this example in Section VI.

C. Relation Between $\Lambda^*(\cdot)$ and $I(\cdot)$

The two rate functions introduced above, $\Lambda^*(\cdot)$ and $I(\cdot)$, characterize the large deviation principles for the mean transmission time and average service rate, respectively. Since the processes $\{T_q\}$ and $\{D_s\}$ are closely related, one can presume that their governing rate functions are somehow linked. A key insight in understanding this relation is to realize that the following events are equivalent: for any positive integers m and n

$$\{T_1 + \cdots + T_m > n\} = \{D_1 + \cdots + D_n < m\}. \tag{19}$$

In words, the first event occurs whenever more than n attempts are required to successfully deliver m packets, while the second event states that fewer than m packet transmissions have been successful within the first n attempts. Using this relationship and scaling arguments, one can establish our next proposition which substantiates the existence of a strong connection between the two rate functions.

Proposition 3: If the rate functions $\Lambda^*(\cdot)$ and $I(\cdot)$ are finite in the open intervals $(1, \infty)$ and $(0, 1)$, respectively, then they satisfy

$$I(x) = x\Lambda^*\left(\frac{1}{x}\right)$$

for $x \in (0, 1)$.

Proof: See Appendix F. ■

V. Performance Evaluation

Thus far, we have devoted much attention to developing a thorough understanding of H_0 and, in particular, its generating function. In this section, we apply the results of Theorem 2 and we derive a number of pertinent performance criteria with practical significance.

First, recall that $[\![z^t]\!]\, G_{H_0}(z) = \Pr(H_0 = t)$. Accordingly, the probability that the queue fails to drain within τ time units is equal to

$$\Pr(H_0 > \tau) = 1 - \sum_{t=0}^{\lfloor\tau\rfloor} [\![z^t]\!]\, G_{H_0}(z).$$

Moreover, the average time required to empty the queue is obtained by differentiating the moment generating function of H_0 and then taking the limit as z approaches one

$$\mathrm{E}[H_0] = \lim_{z\uparrow 1} \frac{d}{dz} G_{H_0}(z).$$

Alternatively, using Chernoff inequalities, it is possible to upper bound the probability of a deviation event in a computationally efficient manner. The equation

$$\Pr(H_0 > \tau) \le e^{-\lambda\tau}\mathrm{E}\left[e^{\lambda H_0}\right] = e^{-\lambda\tau} G_{H_0}\left(e^\lambda\right)$$

holds for any $\lambda > 0$. The optimal bound derived from this collection of inequalities is sometimes expressed in logarithmic form

$$\log\Pr(H_0 > \tau) \le -\sup_{\lambda>0}\left\{\lambda\tau - \log\left(G_{H_0}\left(e^\lambda\right)\right)\right\}.$$

The large deviation principle on H_0 derived in Section IV confirms that, under mild conditions, this latter bound is asymptotically tight.

It may be instructive to stress that H_0, the first-passage time introduced in (7), is defined in terms of codeword transmission attempts. That is, H_0 represents the cumulative number of codewords sent by the source until the queue empties out completely. Such a metric poses no issue when comparing systems of identical block lengths. However, when assessing the performance of candidate implementations with different block lengths, a more careful interpretation of the results becomes necessary. This subtlety arises because of the mismatch in indexing between the evolution of the queue and the number of channel uses. For a fair evaluation of potential candidates, hitting times should be scaled to portray their evolution according to a common clock, that of the channel process.

Define random variable $\tilde{H}_0$ by

$$\tilde{H}_0 = NH_0,$$

where N designates the block length associated with the underlying implementation. Then, $\tilde{H}_0$ denotes the number of channel uses necessary to empty out the queue, and it can therefore be employed to provide a uniform measure of performance. While it is straightforward to extend our performance criteria to $\tilde{H}_0$ through the relation

$$\Pr(H_0 > \tau) = \Pr\left(\tilde{H}_0 > \frac{\tau}{N}\right),$$

it is essential to apply this transformation when comparing systems with different block lengths.

A similar scaling is needed when comparing the large deviations of systems with different parameters. A proper scaling for the fair comparison of mean sojourn times can be expressed in terms of channel uses per information bit

$$\tilde{Y}_\ell = \frac{1}{\ell} N H_0^{(\lceil \ell/K \rceil)}.$$

This leads to the following asymptotic regime:

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

where $\tau > \mathrm{E}\left[\tilde{Y}_\infty\right]$. Likewise, to account for discrepancies in design parameters, the empirical mean service can be expressed in terms of decoded bits per channel use

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

The ensuing asymptotic regime becomes

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

where $\eta < \mathrm{E}\left[\tilde{Z}_\infty\right]$. Collectively, these various modifications enables the comparison of competing implementations with different values for K and N.

Another concern that comes into play when optimizing over block length is the impact of the initial state of the system. If the number of bits at the source is fixed at time zero, the scope of the optimal solution may be very narrow. This is a situation akin to overfitting in statistical modeling. To provide a more robust characterization with widely applicable results and guidelines, it may be beneficial to assume that the number of bits in the queue at the onset of the transmission process is random, with a prescribed representative distribution. In our numerical study, we circumvent some of these difficulties by assuming that the block length is fixed and the initial queue length is random. The specifics of our investigation are detailed below.

VI. Numerical Analysis

In this section, we apply the methodology developed above to an illustrative example. Physical parameters are selected to resemble an implementation of the global system for mobile communications (GSM). Specifically, the block length is fixed at $N = 114$. The information content per codeword K is a parameter to be optimized. We model the wireless connection as a Gilbert–Elliott erasure channel, and we denote its transition probability matrix as

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

For simplicity, we assume that $\varepsilon_1 = 1$ and $\varepsilon_2 = 0$. The probability of a bit erasure is set at 20%, which entails

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

For this elementary model, channel memory can be expressed unambiguously through the decay factor $(1 - b_{12} - b_{21})$, which is determined by the spectrum of the matrix. A decay factor equal to zero is equivalent to a memoryless channel, while correlation increases as $(1 - b_{12} - b_{21})$ approaches one. Except where specified otherwise, we employ a decay factor equal to 0.9 in our numerical results.

We assume that L, the number of information bits contained at the source at time zero, is a random variable possessing a Gamma distribution with mean 2000 and standard deviation 100. Randomizing the number of bits at the source partly alleviates the idiosyncratic effects associated with partitioning the queue content into segments of K bits. For a source buffer with ℓ information bits, the number of segments to be delivered is $\lceil \ell/K \rceil$ and, as such, a one-bit variation in ℓ can result in having an additional message to send. Imposing a random distribution on the number of information bits at the source leads to a probability distribution on $M = \lceil L/K \rceil$. This, in turn, yields smoother results.

Figs. 4 and 5 present the mean and variance of the first-passage times for the ARQ and hybrid ARQ schemes as functions of the number of information bits per codeword. Varying the code rate affects both the expected value of the first-passage time and its variance. A low code rate offers more protection against erasures and, accordingly, the resulting distribution of the hitting time to an empty queue is very narrow. Increasing the code rate initially reduces the mean first-passage time, as every successful decoding attempt reveals more information bits. However, a higher code rate also raises the probability of decoding failure. Eventually, as the code rate is pushed further, decoding failures start to hamper the draining process

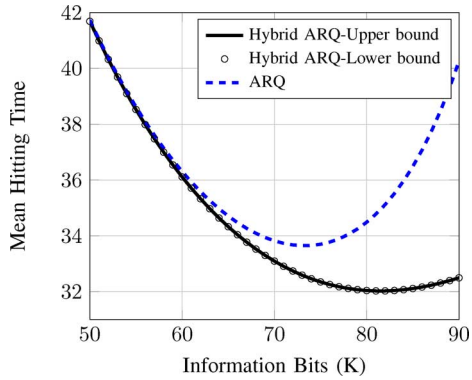

Fig. 4. This figure shows mean first-passage times as functions of K. The block length employed in all cases is $N = 114$. The underlying Gilbert–Elliott channel produces erasures with probability 0.20, and it possesses a dominant decay factor of $(1 - b_{12} - b_{21}) = 0.9$. The expected number of bits at the source at time zero is 2000. The upper and lower bounds for the hybrid ARQ scheme with a depth of $a = 3$ are indistinguishable.

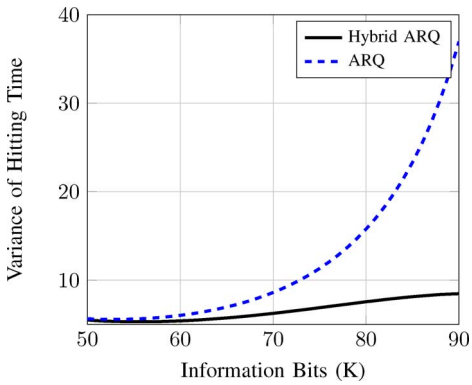

Fig. 5. This figure displays variances of the first-passage times to an empty queue as functions of K. The parameters used in this numerical study are the same as those featured in Fig. 4. The variance for the hybrid ARQ scheme is calculated with the upper bound $\hat{T}$.

and the mean first-passage time grows due to excessive repetition requests. This effect is much more pronounced for standard ARQ.

The penalty in using a high code rate is less severe for the hybrid ARQ scheme because the failure recovery mechanism, which is based on incremental redundancy, adapts gracefully to channel conditions in this latter case. For instance, when K is very close to N, decoding under standard ARQ will fail nearly every time. Contrastingly, the effective code rate drops rapidly with decoding failures under hybrid ARQ. The robust profile of hybrid ARQ is a key property that underlies the popularity of this paradigm in practical systems. In the current example, the upper and lower bounds derived for $\mathrm{E}[H_0]$ under the hybrid ARQ scheme are essentially indistinguishable, hinting at the fact that decoding failures are nearly nonexistent once three blocks are received.

Perhaps not too surprisingly, our numerical investigation suggests that the optimal code rate is somewhat impervious to initial queue conditions. To examine the effects of the initial queue length, we employ the channel parameters described above and we modify the distribution on L. For Gamma distributions with means $\mathrm{E}[L] \in \{500, 1000, 2000, 3000\}$ and standard deviation

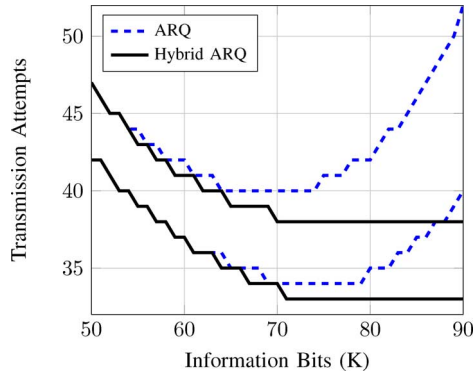

Fig. 6. Crossings of the cumulative distribution function $F_{H_0}(\cdot)$ offer conservative figures of merit for the operation of the queuing system. In this example, the lines correspond to thresholds $p \in \{0.45, 0.95\}$.

100, the optimal value of K in terms of mean first-passage time is consistently equal to 73 for standard ARQ and it remains fixed at 81 for the hybrid variant.

Using the methodology established thus far, it is possible to consider additional performance criteria. For instance, we can analyze the crossings of the cumulative distribution function

$$h_p = \min_t \{t | \Pr(H_0 \leq t) \geq p\}.$$

Fig. 6 plots the number of transmission attempts associated with threshold values $p \in \{0.45, 0.95\}$. We observe that the optimal value of K decreases slightly when the crossing threshold p approaches one. In other words, when focusing on worst case behavior, the system tends to favor a more conservative setting with extra protection against erasures. This phenomenon offers another perspective on the tradeoff between expected behavior and its variations.

Next, we turn to the large deviations techniques developed in Section IV. As a reference, we consider a voice stream application. In GSM, each speech frame of length 20 ms is encoded into a data segment of length 228. The underlying physical layer has the ability to transmit one symbol every 40 μs. If we approximate the maximum delay tolerance for one-way voice traffic to be 40 ms [38, p. 70], then this requires 228 bits to be transmitted within roughly 1000 channel uses. This constraint, in turn, necessitates a nominal rate on the order of 0.23 bits per channel use for link reliability. We adopt this figure as a rough estimate for the needs of a voice stream in our numerical study.

The maximum throughput that can be supported over the Gilbert–Elliott channel in our example is slightly above 0.5 bits per channel use. Recall that threshold η represents a minimum target requirement on the number of information bits per channel use that can be successfully decoded at the destination, in an asymptotic regime. When $\eta < 0.5$, there exist values of K for which the rate function $\frac{1}{N} I\left(\frac{N}{K}\eta\right)$ is strictly positive; this can be seen in Fig. 7. These curves can be used to characterize the tension between quantization and failures to deliver media properly. A high-quality stream, with a large η, will offer an enhanced viewer experience when transmitted adequately, but will necessarily be more prone to interruptions and failures, as exposed through the rate functions. A low-bandwidth, low-quality stream on the other hand offers a better delivery

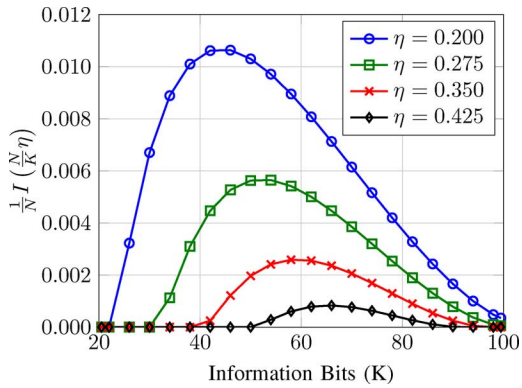

Fig. 7. This figure plots good rate functions governing large deviations in the empirical mean service as functions of K, the number of information bits per codeword. Given throughput threshold η, the optimal value of K is the argument corresponding to the apex of the function.

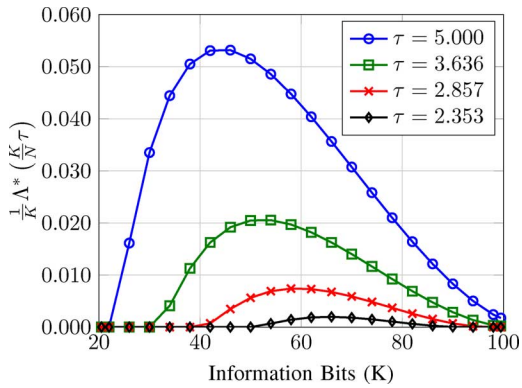

Fig. 8. This figure shows good rate functions governing large deviations in the mean sojourn time as functions of K. The optimum code rate depends heavily on the deviation threshold of the mean sojourn time.

profile with a smaller probability of failure. However, the quality of the playback may not be satisfactory to the end user. A proper selection of parameters for an adequate overall user experience can be made through the rate functions of Fig. 7.

Once η is picked, the corresponding curve displays performance as a function of K. For low code rates, the maximum achievable throughput is less than the service requirement, and hence, the rate function governing large deviations is zero. At high code rates, performance is limited by the rise in the probability of decoding failure. The system must then find the right balance between the frequency of failures and the payoff of a decoding success in terms of information bits. The optimal value of K for a specific threshold η is given by the apex of its curve

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

It is interesting to note how conservative the optimal code rate becomes when the target service requirement is reduced.

The second type of rate functions introduced in Section IV characterizes large deviations in the mean sojourn times, as shown in Fig. 8. These curves can be employed to tradeoff playback quality and buffering times for streaming media. More specifically, τ represents a limitation on the average number of channel uses employed to transmit one bit of information. Of course, when a high-quality rendering is selected, the system must deliver a larger amount of data within the buffering window and, hence, the probability of delay violation becomes greater. In this case, the optimal value of K becomes

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

The behavior of the system in terms of average sojourn time is closely related to the empirical mean service, holding a reciprocal relation. We emphasize that the optimal code rates are equal, namely $K_Z^*(\eta) = K_Y^*(\tau)$ whenever $\tau = \eta^{-1}$. This is due to the relation between $I(\cdot)$ and $\Lambda^*(\cdot)$ described in Section IV-C.

The last aspect of this system we wish to explore is the potential impact of channel memory and correlation among successive channel uses. As before, we keep the probability of a bit erasure at 20%. However, we vary the decay factor of the channel, $(1 - b_{12} - b_{21})$, from zero to one. Once again, we assess performance using the mean first-passage time to an empty queue. When the channel is memoryless, the optimal value for K is 81. As correlation increases, more protection against erasures is beneficial and the optimal value of K decreases moderately. This enables the system to compensate for short sequences of erasures. Still, as correlation strengthens, it becomes difficult to correct longer strings of erasures. When this happens, the penalty of a smaller payoff produced by a low rate code begins to dominate. In other words, attempting to recover every packet starts to be ineffective. Rather, the code rate must be selected to transmit more information bits when the channel is favorable. As $(1 - b_{12} - b_{21})$ approaches one, the optimal value of K/N tends to one as well. In the limit, the channel behaves much like a packet erasure model: send as many bits as possible when the channel is good and ask for retransmissions whenever the message is corrupted. The data points that provide a basis for these findings are summarized in Table I.

TABLE I
OPTIMAL NUMBER OF INFORMATION BITS PER CODEWORD AS A FUNCTION OF CHANNEL MEMORY FACTOR $1 - b_{12} - b_{21}$

Channel Memory	Optimal Value of K		Mean First-Passage Time $\mathrm{E}[H_0]$		Crossing $h_{0.95}$	
	ARQ	HARQ	ARQ	HARQ	ARQ	HARQ
0	81	81	26.92	25.90	30	30
0.5	77	78	28.87	27.79	32	32
0.9	73	81	34.65	32.03	40	38
0.95	77	96	36.68	31.75	44	38
0.98	95	107	35.21	28.62	45	36

VII. CONCLUSION

This paper presents a methodology for the analysis and the design of digital communication systems that operate over channels with memory. The proposed approach is based on the time elapsed between the onset of the communication process and its termination. Results also extend to the asymptotic decay rates of mean service and mean sojourn time. Emphasis is on the selection of code rate for protection against erasures. We provide a simple mathematical characterization of the first-passage time to an empty queue and the large deviations on the mean service and mean transmission time, along with a computationally

efficient means to compare the performance of various implementation candidates.

The properties of coded systems are explored through a numerical study. Optimal code rates appear robust to initial buffer conditions at the transmitter. That is, the number of information bits to be sent from the source to the destination does not significantly affect the optimal operating point of the encoder. Optimal operation is achieved with very similar K values for mean first-passage times and various crossings of the cumulative distribution function.

For both mean service rate 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, delay tolerant applications may be able to use a higher rate on the same physical channel. This phenomenon is closely related to the concept of effective capacity.

Finally, the optimal code rate depends heavily on channel memory. This suggest that, for systems with fixed block lengths, the channel parameters should be estimated and fed back to the encoder for optimal operation. This naturally leads to adaptive strategies and possibly state-aware encoding schemes at the source.

Appendix A
Proof of Theorem 1

We begin this proof by introducing a convenient notation for abstract sequences. Let $\{a_s\}$ be a discrete-time sequence and assume that r and t are two integers with $r < t$. We use a_r^t to denote the subsequence $a_r, a_{r+1}, \ldots, a_t$.

Suppose $u_t = (i_t, q_t) \in \mathcal{C} \times \mathbb{N}_0$ for every $t \geq 0$. Since $\{U_s\}_{s\geq 0}$ is a discrete-time stochastic process whose elements take on values in a finite set, it suffices to show that

$$\Pr\left(U_{s+1} = u_{s+1} | U_0^s = u_0^s\right) = \Pr\left(U_{s+1} = u_{s+1} | U_s = u_s\right)$$

in order to prove that this process is Markov. In general, the probability on the left-hand side can be expressed as

$$\begin{aligned}&\Pr\left(C_{(s+1)N+1} = i_{s+1} | U_0^s = u_0^s\right) \times \\ &\quad \Pr\left(Q_{s+1} = q_{s+1} | U_0^s = u_0^s, C_{(s+1)N+1} = i_{s+1}\right).\end{aligned}$$

We know that the state of the channel at the onset of codeword $s+1$, labeled $C_{(s+1)N+1}$, is conditionally independent of the subsequence Q_0^s and the channel states $C_1^{(s-1)N+1}$, given C_{sN+1}. Thus, we get

$$\begin{aligned}&\Pr\left(C_{(s+1)N+1} = i_{s+1} | U_0^s = u_0^s\right) \\ &\quad = \Pr\left(C_{(s+1)N+1} = i_{s+1} | C_{sN+1} = i_s\right).\end{aligned}$$

The length of the queue Q_{s+1} at time $s+1$ is either Q_s or $Q_s - 1$, depending on whether a codeword is successfully decoded at time s. For a nonempty queue, this depends solely on the generated codebook and the channel realizations during the transmission cycle of the codeword s. As such, we can write

$$\begin{aligned}&\Pr\left(Q_{s+1} = q_{s+1} | U_0^s = u_0^s, C_{(s+1)N+1} = i_{s+1}\right) \\ &\quad = \Pr\left(Q_{s+1} = q_{s+1} | U_s = u_s, C_{(s+1)N+1} = i_{s+1}\right).\end{aligned}$$

Collecting these two results, we conclude that $\{U_s\}$ possesses the Markov property.

Appendix B
Proof of Proposition 1

Notice that the proposition is trivially true when $e > p$. The only case of interest then corresponds to $e \leq p$. We observe that, through a change in indexing, we can write

$$\prod_{l=0}^{n+e-1}\left(1 - 2^{l-p-n}\right) = \prod_{l=-n}^{e-1}\left(1 - 2^{l-p}\right).$$

As such, we readily see that $P_{\mathrm{f}}(p+n, e+n)$ is monotonically increasing in n. The difference between this function and $P_{\mathrm{f}}(p, e)$ is obtained as follows:

$$\begin{aligned}&P_{\mathrm{f}}(p+n, e+n) - P_{\mathrm{f}}(p, e) \\ &\quad = \prod_{l=0}^{e-1}\left(1 - 2^{l-p}\right) - \prod_{l=0}^{n+e-1}\left(1 - 2^{l-n-p}\right) \\ &\quad = \prod_{l=n}^{n+e-1}\left(1 - 2^{l-n-p}\right) - \prod_{l=0}^{n+e-1}\left(1 - 2^{l-n-p}\right) \\ &\quad = \prod_{l=n}^{n+e-1}\left(1 - 2^{l-n-p}\right)\left(1 - \prod_{l=0}^{n-1}\left(1 - 2^{l-n-p}\right)\right) \\ &\quad \leq 1 - \prod_{l=0}^{n-1}\left(1 - 2^{l-n-p}\right) \overset{(a)}{\leq} \sum_{l=0}^{n-1} 2^{l-n-p} \\ &\quad = \sum_{l=0}^{n-1} 2^{-l-1-p} \leq \sum_{l=0}^{\infty} 2^{-l-1-p} = 2^{-p}.\end{aligned}$$

Step (a) follows from an n-variable version of the inequality $1 - (1-p_1)(1-p_2) \leq p_1 + p_2$ where $0 \leq p_1, p_2 \leq 1$. This concludes the demonstration.

Appendix C
Proof of Lemma 2

When $\lambda < -\log \varrho(\mathbf{K})$, the spectral radius of the matrix $\mathbf{K}e^{\lambda}$ is strictly less than one, and consequently, the matrix $\mathbf{I} - \mathbf{K}e^{\lambda}$ is invertible. The finiteness of $\mathbf{G}_T\left(e^{\lambda}\right)$ immediately follows. We then turn to the alternate case, which we prove by contradiction.

Assume that, for some $\lambda \geq -\log \varrho(\mathbf{K})$, matrix $\mathbf{G}_T\left(e^{\lambda}\right)$ exists over the nonnegative real numbers. Note that this condition implies $\varrho(\mathbf{K}) > 0$. For convenience, we wish to work with the

irreducible normal form of $\mathbf{K}$ [36]. That is, there exists a permutation matrix $\mathbf{P}$ such that

$$\tilde{\mathbf{K}} = \mathbf{P}^T\mathbf{K}\mathbf{P} = \begin{bmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{\Phi}_{12} & \cdots & \boldsymbol{\Phi}_{1h} \\ \mathbf{0} & \boldsymbol{\Psi}_2 & \cdots & \boldsymbol{\Phi}_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Psi}_h \end{bmatrix}$$

in which each $\boldsymbol{\Psi}_i$ is either irreducible or a zero matrix. Of course, this reordering also affects $\mathbf{M}$

$$\tilde{\mathbf{M}} = \mathbf{P}^T\mathbf{M}\mathbf{P}.$$

However, this transformation does not alter the spectrum of $\mathbf{K}$ or $\mathbf{M}$. We note that all the states corresponding to an irreducible $\boldsymbol{\Psi}_i$ belong to a same communicating class, which we denote by $\mathcal{C}_i$. Looking at the block triangular structure of $\tilde{\mathbf{K}}$, we gather that the eigenvalues of $\tilde{\mathbf{K}}$ correspond to the union of the eigenvalues of $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_h$. Thus, there exists an integer j such that $\varrho(\boldsymbol{\Psi}_j) = \varrho(\mathbf{K})$.

Since matrix $\boldsymbol{\Psi}_j$ is nonnegative and irreducible, the Perron–Frobenius theorem applies and there exists an eigenvector $\mathbf{v}$, with positive components, such that

$$\mathbf{v}\boldsymbol{\Psi}_j = \varrho(\boldsymbol{\Psi}_j)\mathbf{v} = \varrho(\mathbf{K})\mathbf{v}.$$

Without loss of generality, we can assume that $\mathbf{v}$ is normalized to one. Let $\mathbf{w}$ be a probability distribution with weight $\mathbf{v}$ over the states associated with $\boldsymbol{\Psi}_j$ and zero elsewhere, i.e.,

$$\mathbf{w} = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{v} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}.$$

Because $\mathbf{v}$ is an eigenvector of $\boldsymbol{\Psi}_j$, we have

$$\mathbf{w}\left(\tilde{\mathbf{K}}e^{\lambda}\right)^t = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \left(\varrho(\mathbf{K})e^{\lambda}\right)^t\mathbf{v} & * & \cdots & * \end{bmatrix}$$

and, correspondingly,

$$\begin{aligned} \mathbf{w}\sum_{t=0}^{\infty}\tilde{\mathbf{K}}^t e^{t\lambda} &= \sum_{t=0}^{\infty}\mathbf{w}\tilde{\mathbf{K}}^t e^{t\lambda} \\ &= \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \sum_{t=0}^{\infty}\left(\varrho(\mathbf{K})e^{\lambda}\right)^t\mathbf{v} & * & \cdots & * \end{bmatrix}. \end{aligned}$$

We note that the multiplicative factor $\sum_{t=0}^{\infty}\left(\varrho(\mathbf{K})e^{\lambda}\right)^t$ is a divergent sum that increases to infinity. In fact, all the components of $\mathbf{w}\sum_{t=0}^{\infty}\tilde{\mathbf{K}}^t e^{t\lambda}$ corresponding to states that are accessible from $\mathcal{C}_j$ must also diverge [36]. Since by assumption the elements of

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

remain finite, we conclude that any state accessible from $\mathcal{C}_j$ must lie in the null space of $\tilde{\mathbf{M}}$. This necessarily means that $\mathbf{w}\tilde{\mathbf{G}}_T\left(e^{\lambda}\right) = \mathbf{0}$, and consequently, $\mathbf{w}\tilde{\mathbf{G}}_T(1) = \mathbf{0}$ because $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{M}}$ are nonnegative matrices. In other words, we have created a valid probability distribution $\mathbf{w}$ for which $\mathbf{w}\tilde{\mathbf{G}}_T(1) = \mathbf{0}$. Equivalently, in the original domain, we can rewrite this equation as $\mathbf{w}\mathbf{P}^T\mathbf{G}_T(1) = \mathbf{0}$. But this equation violates our assumption that T is finite almost surely. We then conclude, by contradiction, that not all entries of $\mathbf{G}_T\left(e^{\lambda}\right)$ are finite when $\lambda \geq -\log\varrho(\mathbf{K})$.

Appendix D
Proof of Corollary 1

As a straightforward application of Lemma 2, we can show that $\varrho(\mathbf{K}) < 1$. By design, we know that T is finite almost surely. Then, from the definition of the matrix generating function $\mathbf{G}_T(z)$ in (9), we gather that

$$\begin{aligned} \left[\mathbf{G}_T(1)\right]_{ij} &= \mathrm{E}\left[\mathbf{1}_{\{C_{NT+1}=j\}}|C_1 = i\right] \\ &= \Pr\left(C_{NT+1} = j|C_1 = i\right). \end{aligned}$$

That is, $\mathbf{G}_T(1)$ is a right stochastic matrix.

Since $\mathbf{K}$ is a substochastic matrix, we already have the relation $\varrho(\mathbf{K}) \leq 1$. We wish to show that, in the current framework, this inequality is strict. Suppose that $\varrho(\mathbf{K}) = 1$. Lemma 2 states that, if $\lambda = -\log\varrho(\mathbf{K}) = 0$; then, not all entries of $\mathbf{G}_T\left(e^0\right) = \mathbf{G}_T(1)$ can be finite. In particular, $\mathbf{G}_T(1)$ cannot be a right stochastic matrix. This leads to an obvious contradiction, which indicates that $\varrho(\mathbf{K}) < 1$, as desired.

Appendix E
Proof of Proposition 2

For the first part of this proof, we assume that $\lambda < -\log\varrho(\mathbf{K})$. The spectral radius of $\mathbf{K}e^{\lambda}$ is then strictly less than one and, as such, $\left(\mathbf{I} - \mathbf{K}e^{\lambda}\right)$ is invertible. This implies that the matrix

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

is well defined over the real numbers. Under Assumption 2, we know that $\mathbf{G}_T(1)$ is an irreducible matrix. This readily implies that $\mathbf{G}_T\left(e^{\lambda}\right)$ is also irreducible. We can therefore apply the Perron–Frobenius theorem [22, Th. 3.1.1], whose asymptotic properties lead directly to $\Lambda(\lambda)$.

For the second case, we suppose that $\lambda \geq -\log\varrho(\mathbf{K})$. By Lemma 2, we know that at least one entry of $\mathbf{G}_T\left(e^{\lambda}\right)$ is equal to infinity. We can use the irreducibility of this matrix to argue that each row in $\left(\mathbf{G}_T\left(e^{\lambda}\right)\right)^k$ has at least one entry that is infinite. Since π_0 is a probability distribution

$$\mathrm{E}\left[e^{\lambda(T_1+\cdots+T_k)}\right] = \pi_0\left(\mathbf{G}_T\left(e^{\lambda}\right)\right)^k\mathbf{1} = \infty.$$

For any $m > k$, we have

$$\begin{aligned} \Lambda_m(m\lambda) = \log\mathrm{E}\left[e^{m\lambda Y_m}\right] &= \log\mathrm{E}\left[e^{\lambda(T_1+\cdots+T_m)}\right] \\ &\geq \log\mathrm{E}\left[e^{\lambda(T_1+\cdots+T_k)}\right] = \infty. \end{aligned}$$

Consequently, whenever $\lambda \geq -\log \varrho(\mathbf{K})$, we get

$$\Lambda(\lambda) = \lim_{m\to\infty} \frac{1}{m}\Lambda_m(m\lambda) = \infty,$$

as desired.

Appendix F
Proof of Proposition 3

For the sake of completeness, we offer a brief proof for Proposition 3. As an initial step for this demonstration, we establish a few key properties. The processes $\{Y_m\}$ and $\{Z_s\}$ converge almost surely, i.e.,

$$Y_m = \frac{1}{m}\sum_{q=1}^{m} T_q \to \bar{T} \text{ a.s.}$$
$$Z_s = \frac{1}{s}\sum_{t=1}^{s} D_t \to \bar{D} \text{ a.s.},$$

where $\bar{T}$ are $\bar{D}$ are constants. Moreover, $\bar{T}$ and $\bar{D}$ have a reciprocal relation, i.e., $\bar{T} = 1/\bar{D}$.

Recall that process $\{V_s = (C_{(s+1)N+1}, D_s)\}$ is a finite-state Markov chain with irreducible transition probability matrix $\mathbf{\Pi}$. Also, $D_s = f(V_s)$ is a (trivial) bounded function. Then, by the ergodic theorem for Markov chains [20], we have

$$\Pr\left(\lim_{s\to\infty} \frac{1}{s}\sum_{t=1}^{s} D_t = \bar{D}\right) = 1.$$

Let Ω_1 be the subset of Ω defined by

$$\Omega_1 = \left\{\omega : \frac{1}{s}\sum_{t=1}^{s} D_t(\omega) \to \bar{D}\right\}.$$

Clearly, for any $\omega \in \Omega_1$, we necessarily have

$$N(s,\omega) = \sum_{t=1}^{s} D_t(\omega) \to \infty.$$

Consider the empirical average defined by

$$\frac{1}{m}\sum_{q=1}^{m} T_q. \tag{20}$$

We wish to show that this sequence converges almost surely to $1/\bar{D}$ as m increases to infinity. For any $\omega \in \Omega_1$, we have

$$\sum_{q=1}^{N(s,\omega)} T_q(\omega) \leq s \leq \sum_{q=1}^{N(s,\omega)+1} T_q(\omega).$$

As such, we get the inequality

$$\frac{1}{N(s,\omega)}\sum_{q=1}^{N(s,\omega)} T_q(\omega) \leq \frac{s}{N(s,\omega)} \to \frac{1}{\bar{D}}.$$

In a similar fashion, we obtain

$$\frac{1}{N(s,\omega)+1}\sum_{q=1}^{N(s,\omega)+1} T_q(\omega) \geq \frac{s}{N(s,\omega)+1} = \frac{N(s,\omega)}{N(s,\omega)+1}\frac{s}{N(s,\omega)} \to \frac{1}{\bar{D}}.$$

It follows that, for any $\omega \in \Omega_1$, we get

$$\frac{1}{N(s,\omega)}\sum_{q=1}^{N(s,\omega)} T_q(\omega) \to \frac{1}{\bar{D}}. \tag{21}$$

To complete the proof, we must connect this result to our original sequence (20). We emphasize that, for any $\omega \in \Omega_1$ and for any $m \in \mathbb{N}$, there exists s such that $N(s,\omega) = m$ because $N(s,\omega)$ increases by at most one at every step. It follows that (20) is a subsequence of convergent sequence (21). They must then share the same limit. Collecting these results, we gather that

$$\Pr\left(\lim_{m\to\infty} \frac{1}{m}\sum_{q=1}^{m} T_q = \frac{1}{\bar{D}}\right) = 1.$$

As a side note, it is possible to show that

$$\bar{D} = \mathrm{E}_{\pi_D}[D_t] = \pi_D \mathbf{M}\mathbf{1}$$
$$\bar{T} = \mathrm{E}_{\pi_T}[T_q] = \pi_T\left[\lim_{\lambda\uparrow 0}\frac{d}{d\lambda}\mathbf{G}_T\left(e^{\lambda}\right)\right]\mathbf{1},$$

where $\frac{d}{d\lambda}\mathbf{G}_T(e^{\lambda})$ denotes the entrywise derivative. Above, π_D and π_T represent the invariant distributions of the channel and the stochastic matrix $\mathbf{G}_T(1)$, respectively.

Our strategy to finish this proof is to establish the claimed result for rational numbers, and then invoke continuity to get a full characterization. From our hypotheses, we know that the rate functions $\Lambda^*(\cdot)$ and $I(\cdot)$ are finite in the open intervals $(1,\infty)$ and $(0,1)$, respectively. We note that these functions are also convex over these intervals and, hence, continuous. Let $r = p/q$, where $p, q \in \mathbb{N}$, be a rational number less than one. Recall that $I(\cdot)$ is convex and, therefore, continuous over $(0,1)$. Then, for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$-I(r) - \epsilon \leq \liminf_{n\to\infty}\frac{1}{np}\log\Pr\left(Z_{np} \in (r-\delta, r+\delta)\right)$$
$$\leq \limsup_{n\to\infty}\frac{1}{np}\log\Pr\left(Z_{np} \in (r-\delta, r+\delta)\right) \leq -I(r) + \epsilon.$$

Taking the limit as $\delta \to 0$, we get

$$\lim_{\delta\to 0}\liminf_{n\to\infty}\frac{1}{np}\log\Pr\left(Z_{np} \in (r-\delta, r+\delta)\right)$$
$$= \lim_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{np}\log\Pr\left(Z_{np} \in (r-\delta, r+\delta)\right) = -I(r).$$

A similar argument applies to $\{Y_m\}$. Noting that $q/p \in (1,\infty)$, we gather that $\Lambda^*(\cdot)$ is continuous in a neighborhood of $1/r$. Then, for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\begin{aligned}
&-\Lambda^*\left(\frac{1}{r}\right) - \epsilon \\
&\quad \le \liminf_{n\to\infty} \frac{1}{nq} \log \Pr\left(Y_{nq} \in \left(\frac{1}{r}-\delta, \frac{1}{r}+\delta\right)\right) \\
&\quad \le \limsup_{n\to\infty} \frac{1}{nq} \log \Pr\left(Y_{nq} \in \left(\frac{1}{r}-\delta, \frac{1}{r}+\delta\right)\right) \\
&\quad \le -\Lambda^*\left(\frac{1}{r}\right) + \epsilon.
\end{aligned}$$

As before, this implies that

$$\begin{aligned}
&\lim_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{nq} \log \Pr\left(Y_{nq} \in \left(\frac{1}{r}-\delta, \frac{1}{r}+\delta\right)\right) \\
&= \lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{nq} \log \Pr\left(Y_{nq} \in \left(\frac{1}{r}-\delta, \frac{1}{r}+\delta\right)\right) \\
&= -\Lambda^*\left(\frac{1}{r}\right).
\end{aligned}$$

We stress that the rate functions $\Lambda^*(\cdot)$ and $I(\cdot)$ vanish at $\bar{T}$ and $\bar{D}$, respectively.

At this point, we need to consider two separate cases. First, suppose $r < \bar{D}$. We know that $I(\cdot)$ is a nonincreasing function over interval $\left[0, \bar{D}\right)$ (see, e.g., [22, Lemma 2.2.5]). Also, in an analogous manner, rate function $\Lambda^*(\cdot)$ is nondecreasing over $\left(\bar{T}, \infty\right)$. Leveraging (19), we can write

$$\Pr\left(\frac{T_1 + \cdots + T_{pn}}{pn} > \frac{q}{p}\right) = \Pr\left(\frac{D_1 + \cdots + D_{qn}}{qn} < \frac{p}{q}\right).$$

By letting n go to infinity, we obtain

$$\inf_{x\in\left[\frac{1}{r},\infty\right)} r\Lambda^*(x) = \inf_{x\in(0,r]} I(x).$$

Using the monotonic properties of these rate functions over the prescribed intervals, we get

$$r\Lambda^*\left(\frac{1}{r}\right) = \inf_{x\in\left[\frac{1}{r},\infty\right)} r\Lambda^*(x) = \inf_{x\in(0,r]} I(x) = I(r),$$

as desired.

For the second case, assume $r > \bar{D}$. Under this constraint, the monotonic properties of the rate functions are reversed. That is, $I(\cdot)$ is nondecreasing over $\left(\bar{D}, 1\right)$ and $\Lambda^*(\cdot)$ is nonincreasing over $\left(0, \bar{T}\right)$. Using these relations and the set equalities

$$\Pr\left(\frac{T_1 + \cdots + T_{pn}}{pn} < \frac{q}{p}\right) = \Pr\left(\frac{D_1 + \cdots + D_{qn}}{qn} > \frac{p}{q}\right),$$

we can write

$$r\Lambda^*\left(\frac{1}{r}\right) = \inf_{x\in\left(0,\frac{1}{r}\right]} r\Lambda^*(x) = \inf_{x\in[r,\infty)} I(x) = I(r).$$

Collecting these results, we deduce that $I(x) = x\Lambda^*\left(\frac{1}{x}\right)$ whenever $x \in \mathbb{Q} \cap (0,1)$. Since the rational numbers are dense in $(0,1)$ and the two rate functions are continuous, this equality must also hold for any real number in $(0,1)$.

REFERENCES

[1] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: Wiley, 1968.

[2] R. Negi and J. M. Cioffi, "Delay-constrained capacity with causal feedback," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2478–2494, Sep. 2002.

[3] W. Turin and M. Zorzi, "Performance analysis of delay-constrained communications over slow Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 801–807, Oct. 2002.

[4] I. Bettesh and S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4115–4141, Sep. 2006.

[5] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1468–1489, Jul. 1999.

[6] S. Kittipiyakul, P. Elia, and T. Javidi, "High-SNR analysis of outage-limited communications with bursty and delay-limited information," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 746–763, Feb. 2009.

[7] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[8] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[9] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

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

[11] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.

[12] X. Wang, Q. Liu, and G. B. Giannakis, "Analyzing and optimizing adaptive modulation coding jointly with ARQ for QoS-guaranteed traffic," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 710–720, Mar. 2007.

[13] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[14] R. Negi and S. Goel, "An information-theoretic approach to queuing in wireless channels with large delay bounds," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2004, pp. 116–122.

[15] S. Goel and R. Negi, "The queued-code in finite-state Markov fading channels with large delay bounds," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 30–34.

[16] R. Fantacci, "Queuing analysis of the selective repeat automatic repeat request protocol wireless packet networks," *IEEE Trans. Veh. Technol.*, vol. 45, no. 2, pp. 258–264, May 1996.

[17] R. E. Azouzi and E. Altman, "A queuing analysis of packet dropping over a wireless link with retransmissions," in *Wireless Personal Communications*. Berlin, Germany: Springer-Verlag, 2003, pp. 321–333.

[18] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communications Networks*. New York, NY: Springer-Verlag, 2001.

[19] W. Wu, A. Arapostathis, and S. Shakkottai, "Optimal power allocation for a time-varying wireless channel under heavy-traffic approximation," *IEEE Trans. Autom. Control*, vol. 51, no. 4, pp. 580–594, Apr. 2006.

[20] J. R. Norris, *Markov Chains*. Cambridge, U.K.: Cambridge Univ. Press, 1998, Cambridge Series in Statistical and Probabilistic Mathematics.

[21] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed. Reading, MA: Addison-Wesley, 1994.

[22] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY: Springer-Verlag, 2009, vol. 38.

[23] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. New York, NY: Wiley-Interscience, 1975.

[24] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, 4th ed. New York, NY: Wiley, 2008, Probability and Statistics.

[25] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, no. 9, pp. 1253–1265, 1960.

[26] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, no. 9, pp. 1977–1997, 1963.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: Wiley-Interscience, 1991.
[28] 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, Feb. 1995.
[29] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.
[30] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels: A survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.
[31] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[32] 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.
[33] R. A. Comroe and D. J. Costello, Jr., "ARQ schemes for data transmission in mobile radio systems," *IEEE Trans. Commun.*, vol. 2, no. 4, pp. 472–481, Jul. 1984.
[34] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1311–1321, Aug. 2004.
[35] L. B. Le, E. Hossain, and M. Zorzi, "Queueing analysis for GBN and SR ARQ protocols under dynamic radio link adaptation with non-zero feedback delay," *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3418–3428, Sep. 2007.
[36] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[37] P. Lancaster and M. Tismenetsky, *The Theory of Matrices: With Applications*, 2nd ed. New York, NY: Academic, 1985.
[38] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005.

Santhosh Kumar (S'13) is currently pursuing his Ph.D. in the Department of Electrical and Computer Engineering at Texas A&M University, College Station, TX. His research interests include information and coding theory, wireless communications and statistical inference.

Jean-François Chamberland (S'98–M'04–SM'09) received the Ph.D. degree in 2004 from the University of Illinois at Urbana-Champaign, the M.S. degree in 2000 from Cornell University, Ithaca, NY, and the B.Eng. degree in 1998 from McGill University, Montreal, Canada, all in electrical engineering. He joined Texas A&M University in 2004, where he is currently an associate professor in the Department of Electrical and Computer Engineering. His research interests include communication systems, queueing theory, detection and estimation, and statistical signal processing. In 2006, he was the recipient of a Young Author Best Paper Award from the IEEE Signal Processing Society. He also received a CAREER Award from the National Science Foundation in 2008.

Henry D. Pfister (S'99–M'03–SM'09) received his Ph.D. in electrical engineering from the University of California, San Diego, in 2003 and joined the Department of Electrical and Computer Engineering at Texas A&M University in 2006, where he is currently an Associate Professor. Prior to that, he spent two years in R&D at Qualcomm, Inc. and one year as a post-doc at EPFL. He received the NSF Career Award in 2008 and was a coauthor of the 2007 IEEE COMSOC best paper in Signal Processing and Coding for Data Storage. His current research interests include information theory, channel coding, and iterative information processing with applications in wireless communications, data storage, and signal processing.