Code-Rate Selection, Queueing Behavior, and the Correlated Erasure Channel

Parimal Parag, *Student Member, IEEE*, Jean-Francois Chamberland, *Senior Member, IEEE*, Henry D. Pfister, *Senior Member, IEEE*, and Krishna Narayanan, *Senior Member, IEEE*

***Abstract*—This paper considers the relationship between code-rate selection and queueing performance for communication systems subject to time-varying channel conditions. While error-correcting codes offer protection against channel uncertainties, there exists a natural tradeoff between the enhanced protection of low-rate codes and the rate penalty imposed by additional redundancy. In the limiting regime where codewords are asymptotically long, this tradeoff is well understood and characterized by the Shannon capacity. However, for delay-sensitive communication systems and finite block lengths, a complete characterization of this tradeoff is not fully developed. This paper offers a new perspective on the queueing performance of communication systems with finite block lengths operating over correlated erasure channels. A rigorous framework that links code rate to overall system performance for random codes is presented. Guidelines for code-rate selection in delay-sensitive systems are identified. These findings are supported by a numerical study.**

***Index Terms*—Block codes, channel models, communication systems, data communication, queueing analysis, telecommunication buffers, wireless communication.**

I. Introduction

THE transmission of digital information over noisy channels has become commonplace in modern communication systems. The dependability of contemporary data links is due, partly, to the many successes of information theory and error-control coding [1]. In particular, the reliable transmission of digital information is possible at rates approaching the Shannon capacity using asymptotically long codewords [2]. Indeed, many notable communication systems employ long codewords to provide high throughput and low error probabilities [3].

A scenario where the many insights offered by classical information theory do not apply directly is the broad area of delay-constrained communications [4]. Real-time traffic and live interactive sessions are very sensitive to latency. Long codewords are not particularly well suited for real-time applications because they entail lengthy encoding/decoding delays. In such instances, alternative engineering methods, including power control, automatic repeat-request, scheduling, and feedback, can be leveraged to establish rapid end-to-end connections [5], [6]. Yet, delay considerations often force systems to operate well below their Shannon limits [7].

Several articles in information theory are focused on the tradeoff between throughput and delay. Coding performance as a function of delay has been assessed in the information theory literature using the reliability function [2]. This performance criterion identifies the error exponent of a code family as a function of data rate. The notion of reliability function can be extended to variable-length codes in the presence of feedback, leading to the famous Burnashev error exponent [8]–[10]. While significant, these results remain asymptotic in nature and do not capture the queueing aspect of many communication systems.

Pertinent alternative approaches include effective capacity [11], [12], outage capacity [13], [14], average delay characterizations [15], fluid analysis [16], and heavy-traffic limits [17]. While these contributions also provide insights about the design of delay-sensitive systems, many such articles make idealized assumptions about the behavior of coded transmissions. For instance, some authors adopt the notion of instantaneous capacity: individual data blocks are assumed, implicitly or explicitly, to possess enough degrees of freedom to support sophisticated coding schemes and thereby approach Shannon capacity within every time slot. Perhaps reasonable for long codewords, such assumptions become more of a concern for short data blocks. This is especially problematic for channels with memory, where correlation over time promotes deviations from expected behavior.

For a delay-constrained communication system that utilizes short codewords, two competing goals affect the selection of an error-correcting code. A low-rate code will, in general, result in a small probability of decoding failure, whereas the same system with a high-rate code is more prone to errors. Still, the successful decoding of a codeword associated with a higher rate code leads to the transmission of a larger number of information bits. This tension has already been exposed for communication systems with automatic repeat requests in the context of block-fading channels [18]. For instance, the throughput-maximizing scheme for a system with a short block code may only provision limited resources against channel uncertainties. Indeed, the optimum probability of decoding failure at the block level can remain quite large.

Many previous inquiries in this area adopt a higher layer viewpoint, using rudimentary models for the physical layer; others embrace a channel-coding perspective, intentionally

Manuscript received January 12, 2011; revised August 26, 2011 and March 30, 2012; accepted May 08, 2012. Date of publication September 17, 2012; date of current version December 19, 2012. This work was supported by the National Science Foundation under Grants 0747363, 0747470, and 0830696. This paper was presented in part at the 2010 IEEE Information Theory Workshop, Cairo, Egypt, and in part at the 2010 IEEE International Symposium on Information Theory.

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

Communicated by R. Berry, Associate Editor for Communication Networks.

Digital Object Identifier 10.1109/TIT.2012.2216501

disregarding queueing aspects of the system. Herein, we seek to bridge the gap between these extremes to address an important question: What is the optimal code rate for a particular implementation and traffic profile? Our approach in obtaining an answer to this question differs from established work in that we strive to provide exact solutions. To facilitate the type of queueing analysis we wish to carry, we make the following assumptions. The packet arrival process at the transmitter is Bernoulli, and the packet length has a geometric distribution in bits. The communication medium is a bit-erasure channel with memory. Random codes, with maximum-likelihood decoding, are employed to protect the transmitted data against erasures.

Collectively, these assumptions are sufficient to conduct a rigorous analysis of the probability of block decoding failure at the receiver, which leads to a complete characterization of the ensuing queueing behavior at the source. Implicit in our system model is the ability to acknowledge the reception of packets through instantaneous feedback. We emphasize that model components are selected with the intent to keep analysis manageable. The focus is on developing tools and techniques that can be used to bridge communication, coding, and queueing. Still, our framework admits several extensions beyond the formulation presented in this paper; some of these extensions are discussed alongside the main results wherever appropriate.

The remainder of this paper is organized as follows. The system model is introduced in Section II, with the probability of block decoding failure being derived in Section II-C. Packet arrivals and departures form the main topic of Section III. Together, they dictate the queueing behavior of the packetized system, which is analyzed in Section IV. Numerical results are contained in Section V. Finally, new insights, concluding remarks and avenues of future research, are discussed in Section VI.

II. System Model

We initiate our exposition of the system we wish to study with a description of the underlying communication channel. Bits are sent from a source to a destination over a Gilbert–Elliott erasure channel. The channel can be in one of two states, which we denote by integers $\{1, 2\}$. In state 1, every transmitted bit is erased with probability ε_1 independently of other bits. Similarly, in state 2, every bit is lost with probability ε_2. Throughout, we assume that $\varepsilon_2 \leq \varepsilon_1 \leq 1$. Transitions between channel states occur according to a Markov process. The probability of jumping to state 2 given that the Markov chain is currently in state 1 is denoted by α. The reverse transition probability from state 2 to state 1 is written as β. The parameters of this Markov chain can be expressed in the form of a transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}. \tag{1}$$

A graphical interpretation of this communication channel appears in Fig. 1. We note that the methodology adopted in this paper admits a larger number of channel states and can be applied to more intricate physical links. A differentiating aspect of the Gilbert–Elliott channel is that it represents the simplest nontrivial instance of a finite-state channel with memory, which leads to a more accessible treatment of the problem. Markov models have been employed to capture the behavior of communication channels in the past, and several studies point to methods of selecting parameters to best match the profiles of communication links at the physical layer [19], [20]. In our framework, the marginal distribution of fades is determined by the stationary distribution of the channel, whereas correlation over time is captured through the spectral gap of the transition probability matrix. At this point, we leave the channel parameters in an abstract form, seeking general solutions.

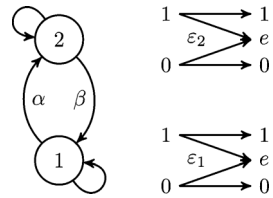

Fig. 1. Gilbert–Elliott bit erasure channel is employed to model the operation of a communication link with memory. This model captures both the uncertainty associated with transmitting bits over a noisy channel and the correlation over time typical of several communication channels.

The channel state at instant n is a random variable, which we label C_n. With this notation, one can write the progression of the Markov chain over time as $\{C_n : n \in \mathbb{N}\}$. Finding the conditional probability $\Pr(C_{n+1} = d | C_n = c)$, where $c, d \in \{1, 2\}$, amounts to selecting an entry in $\mathbf{P}$. Likewise, $\Pr(C_{n+N} = d | C_n = c)$ corresponds to an entry in $\mathbf{P}^N$, where

$$\begin{aligned} \mathbf{P}^N &= \frac{1}{\alpha+\beta} \begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^N \end{bmatrix} \begin{bmatrix} \beta & \alpha \\ 1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta+\alpha(1-\alpha-\beta)^N}{\alpha+\beta} & \frac{\alpha-\alpha(1-\alpha-\beta)^N}{\alpha+\beta} \\ \frac{\beta-\beta(1-\alpha-\beta)^N}{\alpha+\beta} & \frac{\alpha+\beta(1-\alpha-\beta)^N}{\alpha+\beta} \end{bmatrix}. \end{aligned}$$

This decomposition shows how this Markov chain converges to its stationary distribution at an exponential rate that depends on the second largest eigenvalue of $\mathbf{P}$, which is $1 - \alpha - \beta$. This quantity can then be employed to quantify channel memory.

A. Segments, Block Length, and Code Rate

To transmit information over the erasure channel, data bits must first be processed and encoded. In our framework, a packet of length L is sectioned into M data segments, each containing K information bits. Packing loss is accounted for implicitly as the last data segment of a packet is zero-padded to K bits. Thus, the number of segments within a packet of length L is equal to $M = \lceil L/K \rceil$. Data are stored in a queue and every segment is encoded separately into a codeword of length N, which is eventually sent over the Gilbert–Elliott erasure channel. The transmission of a codeword, thus, requires N successive uses of the channel. We assume that decoding failures are handled through immediate retransmission of the missing data. The block length N remains fixed throughout; it is determined by system requirements and the availability of physical resources. On the other hand, the size of a data segment K (and, therefore, the code rate $r = K/N$) is a parameter that should be optimized.

B. Distribution of Erasures

A quantity that is of fundamental importance in our analysis is the probability of decoding failure at the destination. An intermediate step in identifying this probability is to derive expressions for the distributions of the number of erasures E within a

codeword. This, in turn, depends on the number of visits to each state within N consecutive realizations of the channel. Specifically, we are interested in conditional probabilities of the form

$$\Pr(E = e, C_{N+1} = d | C_1 = c) \tag{2}$$

where $c, d \in \{1, 2\}$. The generating functions for these conditional probabilities can be derived based on generalizing the entries of $\mathbf{P}$ to the vector space of real polynomials in x

$$\mathbf{P}_x = \begin{bmatrix} (1-\alpha)(1-\varepsilon_1+\varepsilon_1 x) & \alpha(1-\varepsilon_1+\varepsilon_1 x) \\ \beta(1-\varepsilon_2+\varepsilon_2 x) & (1-\beta)(1-\varepsilon_2+\varepsilon_2 x) \end{bmatrix}.$$

Proposition 2.1: Let $[\![x^k]\!]$ be the linear functional that maps a polynomial in x to the coefficient of x^k. The conditional probability $\Pr(E = e, C_{N+1} = d | C_1 = c)$ is given by

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

where $\mathbf{P}_x$ is the matrix defined previously.

Proof: This result can be shown using mathematical induction. Let $E_{i:j}$ denote the number of channel erasures occurring between times i and j, inclusively. By construction, the proposition holds for $N = 1$. As an inductive step, assume that (3) is satisfied for $N = n - 1 > 0$. Then, one can write

$$\begin{aligned}
&\Pr\left(E_{1:n} = e, C_{n+1} = c_{n+1} | C_1 = c_1\right) \\
&= \sum_{c_n \in \{1,2\}} \Pr(E_{1:n} = e, C_{n+1} = c_{n+1}, C_n = c_n | C_1 = c_1) \\
&= \sum_{c_n \in \{1,2\}} \sum_{k \in \{0,1\}} \Pr(E_{n:n} = k, C_{n+1} = c_{n+1} | C_n = c_n) \\
&\quad \times \Pr(E_{1:n-1} = e - k, C_n = c_n | C_1 = c_1) \\
&= \sum_{c_n \in \{1,2\}} \sum_{k \in \{0,1\}} [\![x^k]\!] \left[\mathbf{P}_x\right]_{c_n, c_{n+1}} [\![x^{e-k}]\!] \left[\mathbf{P}_x^{n-1}\right]_{c_1, c_n} \\
&= [\![x^e]\!] \left[\mathbf{P}_x^n\right]_{c_1, c_{n+1}}.
\end{aligned}$$

That is, (3) also holds for $N = n$. Since both the basis and the inductive step have been verified, we conclude that the proposition is true for all integers $N \geq 1$. ■

We note that one can employ this method or alternative combinatorial means to obtain closed-form expressions for the desired conditional probabilities [21], [22].

C. Probability of Decoding Failure

At the onset of every transmission attempt, a new code is created to encode K information bits. The code is defined by a random parity-check matrix $\mathbf{H}$ of size $(N - K) \times N$. The entries of $\mathbf{H}$ are selected independently and uniformly over $\{0, 1\}$. This scheme assumes shared randomness between the source and its destination. Maximum-likelihood decoding is used at the destination to decode the received messages. Consequently, the probability of decoding failure becomes a function only of the number of erasures contained within a block. Once the value of E is known, one can compute the probability of decoding failure using the following result.

Proposition 2.2: The probability of decoding failure, given e erasures within a codeword of length N, is equal to

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

Proof: Conditioned on $E = e$, decoding at the destination will succeed if and only if the submatrix of $\mathbf{H}$ formed by choosing the e erased columns has rank e. Furthermore, the probability that a random $p \times e$ matrix over $\mathbb{F}_2$ has rank e is equal to $\prod_{i=0}^{e-1} \left(1 - 2^{i-p}\right)$ [23, p. 73]. Collecting these two results, we obtain the probability of a successful transmission, which implicitly determines the conditional probability of decoding failure given $E = e$. ■

The unconditioned probability of decoding failure at the destination is equal to

$$P_{\mathrm{f}}(N - K) = \mathbb{E}\left[P_{\mathrm{f}}\left(N - K, E\right)\right]$$

where the distribution of E accounts for all the possible channel realizations within a block. While the probability of decoding failure represents an important performance criterion, it alone does not capture the queueing behavior of the system. Time dependencies among decoding failures may also influence the behavior of the queue at the source. Having introduced a mathematical model for the physical channel, we turn to the description of the arrival and departure processes.

III. Arrival and Departure Processes

Data packets enter the queue according to a discrete-time Bernoulli process whose clock is synchronized with codeword transmission intervals. During every codeword transmission attempt, a new packet arrives at the source with probability γ, independently of other time instants. The number of bits in every data packet is random, with packet sizes forming a sequence of independent and identically distributed random variables. The marginal distribution of a packet size is geometric with parameter ρ. In other words, the probability that a packet contains exactly ℓ bits is given by

$$\Pr(L = \ell) = (1 - \rho)^{\ell - 1} \rho \qquad \ell = 1, 2, \ldots$$

where $\rho \in (0, 1)$. The arrival process and the packet-length distribution have been selected, partly, to facilitate the analysis we wish to carry. In particular, the memoryless property of the geometric distribution and the independence over time of the Bernoulli process make for a tractable characterization of queueing behavior. Adopting an intricate arrival process more in tune with a specific application can easily render analysis intractable. This explains why our arrival process conforms to models commonly found in the queueing literature.

Departures from the queue are governed by the underlying Gilbert–Elliott channel and the selected number of parity bits, $N - K$, of our random code. The probability of decoding failure is increasing in code rate, given a fixed block length N. Still, the

successful decoding of a high-rate codeword leads to the transmission of a larger number of information bits. As mentioned before, these competing considerations create a natural tradeoff between data content and probability of decoding failure. Accordingly, the code rate $r = K/N$, or equivalently the number of information bits per data segment K, is a parameter that should be optimized.

Once the code rate is specified, the number of successfully decoded codewords needed to complete a packet transmission is random with $M = \lceil L/K \rceil$. We note that L being a geometric random variable with parameter ρ implies that M is also geometric with parameter

$$\rho_r = \sum_{\ell=1}^{K} (1-\rho)^{\ell-1}\rho = 1-(1-\rho)^K.$$

Thus, the probability that a data packet requires the successful transmission of exactly m codewords becomes

$$\Pr(M=m) = (1-\rho_r)^{m-1}\,\rho_r \qquad m = 1, 2, \ldots.$$

We emphasize that, in the current setting, the number of coded blocks per data packet M retains the memoryless property.

In our formulation, we assume that K is independent of channel state, which simplifies analysis. When side information is present at the transmitter, one can enhance performance by picking K as a function of the current state. Furthermore, even without explicit state knowledge, it is possible to estimate the channel state through available feedback, i.e., the automatic repeat-request sequence. In the latter scenario, the selection of K as a function of state estimates becomes a partially observable Markov decision process. Such problems can be computationally challenging or intractable, and they often necessitate careful consideration. Accounting for the presence of partial state information at the transmitter is an interesting question that is beyond the scope of this paper; we leave this matter as a possible future endeavor. This completes the description of the communication system under study. We proceed below with the characterization of overall performance.

IV. Queueing Behavior

Packets are stored in the queue upon generation by the source, and they remain in this buffer until the corresponding data segments are decoded successfully at the destination. We assume that there are no packet losses and, as such, the transmit buffer has no hard limit. When discussing the size of the queue at the source, two distinct characterizations are possible. The first option is to keep track of the number of packets contained in the queue. The second choice is to track the amount of data awaiting transmission. Although the latter alternative provides a more accurate representation of buffer occupancy in bits, the former option is closely related to the concept of packet delay and it is simpler to analyze. For these reasons, we elect to define the state of the queue as the number of data packets in the queue, as is customary in classical queueing literature [24]–[26].

Recall that, in the proposed setting, a packet of length L is first subdivided into M data segments. Each segment is encoded separately into a codeword of length N, and the resulting message is subsequently sent over the communication channel. Successful receptions are acknowledged instantaneously through feedback, whereas decoding failures trigger immediate retransmissions of the missing blocks. Upon confirmation of an accurate transfer, a data segment is marked as delivered and transmission of the next data block begins. Though the presence of instantaneous feedback is assumed for mathematical convenience, it may also be approximated in practice using high-speed decoders and high-power reverse-link communication.

For the head packet to depart from the queue, the destination must successfully decode the received message and this codeword must be carrying the final parcel of information pertaining to this head packet. Specifically, a packet composed of L bits will require the successful reception of $\lceil L/K \rceil$ codewords before it is removed from the queue. The length of the queue at the onset of block s is denoted by Q_s. The state of the Gilbert–Elliott channel at this instant is represented by C_{sN+1}. Together, these two quantities form the state of our Markov process, $Y_s = (C_{sN+1}, Q_s)$. We emphasize that this state space is countable, with Y_s belonging to $\{1,2\} \times \mathbb{N}_0$. Furthermore, the Markov chain underlying the evolution of our system possesses a special structure. It forms an instance of a discrete-time *quasi-birth-death process*. Luckily, there are many established techniques to analyze such mathematical objects [27]–[29].

Our next step is to examine the transition probabilities of this augmented Markov chain. The probability of jumping from Y_s to Y_{s+1} is given by

$$\begin{aligned}\Pr(Y_{s+1} &= (d, q_{s+1}) | Y_s = (c, q_s)) \\ &= \sum_{e=0}^{N} \Pr\left(Q_{s+1} = q_{s+1} | E = e, Q_s = q_s\right) \\ &\qquad \times \Pr(E = e, C_{(s+1)N+1} = d | C_{sN+1} = c). \quad (4)\end{aligned}$$

We have already introduced, in Section II-B, an efficient methodology to compute conditional probabilities of the form $\Pr(E = e, C_{(s+1)N+1} = d | C_{sN+1} = c)$. Accordingly, it suffices to focus on the other component of each summand, $\Pr\left(Q_{s+1} = q_{s+1} | E = e, Q_s = q_s\right)$, to characterize (4).

We first consider conditional events $\{Q_s = q_s\}$ for which $q_s > 0$. In this case, admissible values for Q_{s+1} are given by $\{q_s - 1, q_s, q_s + 1\}$. Two factors can affect the length of the queue: the arrival of a new data packet and the completion of a packet transmission. The latter occurrence will only take place if a codeword is successfully decoded at the destination and the corresponding data block is the last segment of the head packet. Keeping this fact in mind and using independence between arrivals and departures, we get

$$\begin{aligned}&\Pr\left(Q_{s+1} = q_s + 1 | E = e, Q_s = q_s\right) \\ &\quad = \gamma\big(P_{\mathrm{f}}(N-K, e) + (1 - P_{\mathrm{f}}(N-K, e))(1-\rho_r)\big) \quad (5) \\ &\Pr\left(Q_{s+1} = q_s | E = e, Q_s = q_s\right) \\ &\quad = \gamma\left(1 - P_{\mathrm{f}}(N-K, e)\right)\rho_r + (1-\gamma)\big(P_{\mathrm{f}}(N-K, e) \\ &\quad + (1 - P_{\mathrm{f}}(N-K, e))(1-\rho_r)\big) \quad (6) \\ &\Pr(Q_{s+1} = q_s - 1 | E = e, Q_s = q_s) \\ &\quad = (1-\gamma)\left(1 - P_{\mathrm{f}}(N-K, e)\right)\rho_r. \quad (7)\end{aligned}$$

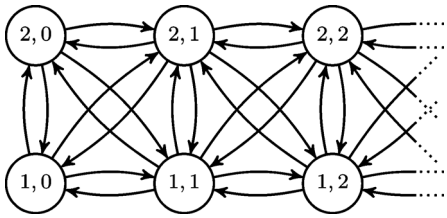

Fig. 2. State space and transition diagram for the aggregate queueing process $\{Y_s\}$; self-transitions are intentionally omitted.

When the queue is empty, no packet can depart. As a result, only two possibilities remain

$$\Pr(Q_{s+1}=1|E=e,Q_s=0)=\gamma \tag{8}$$
$$\Pr(Q_{s+1}=0|E=e,Q_s=0)=1-\gamma. \tag{9}$$

Assembling these results and using (4), we obtain the probability transition matrix of the Markov process $\{Y_s\}$. A graphical representation of possible state transitions appears in Fig. 2.

To proceed with the analysis of our queueing system, a compact representation of the conditional probabilities defined in (4) is apropos. For $q \in \mathbb{N}$ and $c, d \in \{1,2\}$, we introduce the following convenient notation:

$$\mu_{cd}=\Pr(Y_{s+1}=(d,q-1)|Y_s=(c,q)) \tag{10}$$
$$\kappa_{cd}=\Pr(Y_{s+1}=(d,q)|Y_s=(c,q)) \tag{11}$$
$$\lambda_{cd}=\Pr(Y_{s+1}=(d,q+1)|Y_s=(c,q)). \tag{12}$$

Similarly, when the queue is empty, we use

$$\kappa^0_{cd}=\Pr(Y_{s+1}=(d,0)|Y_s=(c,0))$$
$$\lambda^0_{cd}=\Pr(Y_{s+1}=(d,1)|Y_s=(c,0)).$$

Collectively, these labels define the 12 transition probabilities associated with a nonempty queue, and the 8 transition probabilities subject to the nonnegativity constraint at zero.

At last, we are ready to derive the equilibrium distribution of our aggregate system. We note that this system is stable when the mean arrival rate is less than the expected service rate over a codeword transmission period [30], i.e.,

$$\gamma < \rho_r \mathbb{E}\left[1-P_{\mathrm{f}}\left(N-K,E\right)\right].$$

Under this stability condition, Markov chain $\{Y_s\}$ is positive recurrent and possesses a unique stationary distribution [31]. Assuming that the system is stable, let $Y=(C,Q)$ be a random vector with the aforementioned limiting probability distribution

$$\Pr(Y=(c,q))=\lim_{s\to\infty}\Pr(Y_s=(c,q)).$$

We employ the semi-infinite vector π to represent the equilibrium distribution of our system, with

$$\pi(2q+i)=\begin{cases}\Pr(C=1,Q=q) & \text{if } i=0\\ \Pr(C=2,Q=q) & \text{if } i=1.\end{cases}$$

The states $\{(1,q),(2,q)\}$ are known as the qth level of the chain and $\pi_q=[\pi(2q)\ \pi(2q+1)]$ is the stationary distribution associated with this level.

Using this compact notation, one can express the Chapman–Kolmogorov equations for the queued system as $\pi\mathbf{T}=\pi$, where $\mathbf{T}$ denotes the transition probabilities associated with the aggregate Markov chain $\{Y_s\}$. We can represent the transition probability operator $\mathbf{T}$ as a semi-infinite matrix of the form

$$\mathbf{T}=\begin{bmatrix}\mathbf{C}_1 & \mathbf{C}_0 & \mathbf{0} & \mathbf{0} & \cdots\\ \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \cdots\\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \cdots\\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \ddots\end{bmatrix} \tag{13}$$

where the submatrices $\mathbf{C}_1$, $\mathbf{C}_0$, $\mathbf{A}_2$, $\mathbf{A}_1$, and $\mathbf{A}_0$ are 2×2 real matrices. Specifically, we have

$$\mathbf{A}_0=\begin{bmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{bmatrix} \qquad \mathbf{A}_1=\begin{bmatrix}\kappa_{11} & \kappa_{12}\\ \kappa_{21} & \kappa_{22}\end{bmatrix}$$
$$\mathbf{A}_2=\begin{bmatrix}\mu_{11} & \mu_{12}\\ \mu_{21} & \mu_{22}\end{bmatrix}.$$

When the queue is empty, the relevant submatrices become

$$\mathbf{C}_0=\begin{bmatrix}\lambda^0_{11} & \lambda^0_{12}\\ \lambda^0_{21} & \lambda^0_{22}\end{bmatrix} \qquad \mathbf{C}_1=\begin{bmatrix}\kappa^0_{11} & \kappa^0_{12}\\ \kappa^0_{21} & \kappa^0_{22}\end{bmatrix}.$$

When $\mathbf{A}=\mathbf{A}_0+\mathbf{A}_1+\mathbf{A}_2$ is irreducible, this quasi-birth-death process is recurrent if and only if $(v\mathbf{A}_0\mathbf{1})/(v\mathbf{A}_2\mathbf{1})\le 1$, where v is the stationary probability distribution of $\mathbf{A}$ [30]. One possible approach to identify the stationary distribution of the Markov chain $\{Y_s\}$ is to employ spectral representations and ordinary generating functions. This technique is described in Section IV-A. An alternate numerical means for computing the stationary distribution is the matrix-geometric method discussed in Section IV-B. As we will see, both approaches have their advantages and drawbacks.

A. Transform Method

The first approach we present makes use of generating functions [26], [22]. Let $\mathbf{\Pi}(z)$ be the transform vector defined by

$$\mathbf{\Pi}(z)=\sum_{q=0}^{\infty}z^q\pi_q. \tag{14}$$

Theorem 4.1: The invariant distribution of the Markov chain can be derived from the recurrence relation induced by $\mathbf{T}$. Finding the stationary distribution of the augmented Markov chain is equivalent to solving a matrix equation of the form

$$\mathbf{\Pi}(z)\mathbf{D}(z)=\pi_0\left(\mathbf{D}(z)-\mathbf{D}_0(z)\right) \tag{15}$$

where the entries in matrices $\mathbf{D}(z)$ and $\mathbf{D}_0(z)$ are quadratic polynomials

$$\mathbf{D}(z)=z^2\mathbf{A}_0+z\left(\mathbf{A}_1-\mathbf{I}\right)+\mathbf{A}_2 \tag{16}$$
$$\mathbf{D}_0(z)=z^2\mathbf{C}_0+z\left(\mathbf{C}_1-\mathbf{I}\right). \tag{17}$$

The elements of π_0 can be determined from the requirements imposed by stability and normalization.

Proof: We begin this demonstration by writing the balance equations governing the Markov chain $\{Y_s\}$. From the

Chapman–Kolmogorov equations $\pi\mathbf{T} = \pi$ and the form of $\mathbf{T}$ given in (13), we obtain

$$\pi_{q-1}\mathbf{A}_0 + \pi_q(\mathbf{A}_1 - \mathbf{I}) + \pi_{q+1}\mathbf{A}_2 = 0, \quad q \geq 2 \tag{18}$$

$$\pi_0\mathbf{C}_0 + \pi_1(\mathbf{A}_1 - \mathbf{I}) + \pi_2\mathbf{A}_2 = 0 \tag{19}$$

$$\pi_0(\mathbf{C}_1 - \mathbf{I}) + \pi_1\mathbf{A}_2 = 0. \tag{20}$$

Next, we multiply (18) by z^{q+1} and sum over all $q \geq 2$ to get

$$(\mathbf{\Pi}(z) - \pi_0)z^2\mathbf{A}_0 + (\mathbf{\Pi}(z) - \pi_1 z - \pi_0)z(\mathbf{A}_1 - \mathbf{I}) + \left(\mathbf{\Pi}(z) - \pi_2 z^2 - \pi_1 z - \pi_0\right)\mathbf{A}_2 = 0.$$

Leveraging boundary conditions (19) and (20), the aforementioned equation reduces to (15). ■

Using the results of Theorem 4.1, one can write

$$\mathbf{\Pi}(z) = \pi_0\left(\mathbf{I} - \mathbf{D}_0(z)\mathbf{D}^{-1}(z)\right)$$

where $\mathbf{D}^{-1}(z)$ is a matrix whose entries are rational functions of z. Note that one can express the inverse of $\mathbf{D}(z)$ in terms of its adjugate matrix and determinant [32]:

$$\mathbf{D}^{-1}(z) = \frac{\operatorname{adj}\mathbf{D}(z)}{\det\mathbf{D}(z)}.$$

Moreover, the entries of $\mathbf{D}_0(z)\mathbf{D}^{-1}(z)/z$ are rational functions where polynomial numerators have at most degree 3 and the common polynomial denominator $\det\mathbf{D}(z)$ is of degree 4. Through careful inspection, we find that $(z-1)$ is a factor common to all numerator and denominator polynomials. After cancellation, the entries of $\mathbf{D}_0(z)\mathbf{D}^{-1}(z)$ can be expressed as quadratic polynomials over a common cubic polynomial. Using the general formula for the roots of cubic polynomials, it is then possible to carry partial fraction expansion for the entries of $\mathbf{D}_0(z)\mathbf{D}^{-1}(z)$ and thereby obtain an expression for $\mathbf{\Pi}(z)$, which is invertible in closed form. The coefficients of π at level zero are obtained using stability and the fact that

$$\mathbf{\Pi}(1) = \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix}. \tag{21}$$

Although a closed-form parametric solution for the stationary distribution of this system exists and can be obtained using symbolic equation solvers, it is unfortunately too cumbersome to be included in this paper. Still, we emphasize that existence of such a solution provides an efficient means to conduct numerical studies. A downside to the approach outlined previously lies in the fact that it does not scale well with the number of states in the channel, thereby precluding straightforward generalizations to alternate environments. This is due to the difficulty associated with finding the roots of high-degree polynomials. This impediment is addressed in Section IV.B.

B. Matrix-Geometric Method

The Markov chain associated with operator (13) belongs to the class of random processes with repetitive structures. As such, one can apply standard techniques from the rich literature on matrix-analytic methods and quasi-birth-death processes [33]–[36]. The essence of the approach we adopt is to take advantage of the symmetric interactions among different levels of the Markov chain. For $q \geq 2$, the recursive structure of our system is captured by the formula

$$\pi_{q+1}\mathbf{A}_2 = \pi_q(\mathbf{I} - \mathbf{A}_1) - \pi_{q-1}\mathbf{A}_0.$$

In finding a solution to this matrix equation, it seems that the general form of the embedded Markov structure and, specifically, its block partitioning are far more important than the precise values of each submatrix. The stationary distribution of the queue, in matrix-geometric form, is characterized in the following theorem.

Theorem 4.2: Consider a positive recurrent and irreducible Markov chain on a countable state space with transition probabilities given by (13). Let the matrix $\mathbf{U}$ be defined such that the (c,d) entry is the probability that, starting from state $(1,c)$, the Markov chain $\{Y_s\}$ first re-enters level 1 by visiting $(1,d)$ and does so without visiting any state at level 0. The substochastic matrix $\mathbf{U}$ may be computed as the limit, starting from $\mathbf{U}_1 = \mathbf{A}_1$, of the sequence defined by

$$\mathbf{U}_{j+1} = \mathbf{A}_1 + \mathbf{A}_0(\mathbf{I} - \mathbf{U}_j)^{-1}\mathbf{A}_2. \tag{22}$$

Let matrix $\tilde{\mathbf{T}}$ be given by

$$\tilde{\mathbf{T}} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_0 \\ \mathbf{A}_2 & \mathbf{A}_1 + \mathbf{R}\mathbf{A}_2 \end{bmatrix} \tag{23}$$

where $\mathbf{R} = \mathbf{A}_0(\mathbf{I} - \mathbf{U})^{-1}$. Then, $\tilde{\mathbf{T}}$ is a stochastic matrix associated with an irreducible, finite Markov chain. If we denote the invariant distribution associated with $\tilde{\mathbf{T}}$ by $[\tilde{\pi}_0\ \tilde{\pi}_1]$, then the stationary distribution associated with $\mathbf{T}$ can be expressed as

$$\pi_0 = \frac{\tilde{\pi}_0}{(\tilde{\pi}_0 + \tilde{\pi}_1(\mathbf{I} - \mathbf{R})^{-1})\mathbf{1}}$$

$$\pi_q = \frac{\tilde{\pi}_1\mathbf{R}^{q-1}}{(\tilde{\pi}_0 + \tilde{\pi}_1(\mathbf{I} - \mathbf{R})^{-1})\mathbf{1}} \tag{24}$$

where $q \geq 1$.

Proof: See the Appendix. ■

Corollary 4.3: When the appropriate inverse matrices exist, one can write the first two levels of the stationary distribution π associated with (13) as

$$\pi_1 = \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix}\left(\mathbf{A}_2(\mathbf{I} - \mathbf{C}_1)^{-1} + (\mathbf{I} - \mathbf{R})^{-1}\right)^{-1}$$

and $\pi_0 = \pi_1\mathbf{A}_2(\mathbf{I} - \mathbf{C}_1)^{-1}$. The remaining levels are obtained through the recursion $\pi_{q+1} = \pi_q\mathbf{R}$ where $q \geq 1$.

Proof: The coefficients of π_1 can be derived from the channel equilibrium condition

$$\begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} = \sum_{q=0}^{\infty}\pi_q = \pi_0 + \sum_{q=1}^{\infty}\pi_q$$

$$= \pi_1\left(\mathbf{A}_2(\mathbf{I} - \mathbf{C}_1)^{-1} + (\mathbf{I} - \mathbf{R})^{-1}\right).$$

Given that an inverse exists, one can solve for π_1 in terms of the invariant distribution of the channel. From there, the distribution at other levels is obtained in a straightforward manner. ■

In summary, we have presented an algorithmic method to derive the stationary distribution of Y and, concurrently, obtain

the stationary distribution of the queue, $\Pr(Q = q) = \pi_q \mathbf{1}$. It is instructive to note that the matrix $\mathbf{R}$ is closely related to the probability that the number of packets in the queue exceeds a prescribed threshold; it ultimately determines the asymptotic behavior of the complementary cumulative distribution function of the queue.

Corollary 4.4: The decay rate of the complementary cumulative distribution function of the queue satisfies

$$\lim_{\tau \to \infty} \frac{1}{\tau} \log \Pr(Q > \tau) = \log \varrho(\mathbf{R})$$

where $\varrho(\mathbf{R})$ is the spectral radius of $\mathbf{R}$.

Proof: See the Appendix. ■

C. Establishing a Link Between Generating Functions and the Matrix-Geometric Method

We presented two approaches to compute the stationary distribution of the aggregate Markov process. Naturally, these methods must be related. In this section, we explore their connection and we link the generating function procured by first principles to the matrix-geometric method. First, we note that $\mathbf{U}$ satisfies the implicit equation

$$\mathbf{U} = \mathbf{A}_1 + \mathbf{A}_0 (\mathbf{I} - \mathbf{U})^{-1} \mathbf{A}_2. \tag{25}$$

Using the relation $\mathbf{R} = \mathbf{A}_0 (\mathbf{I} - \mathbf{U})^{-1}$ and rearranging terms in (25), we get $\mathbf{A}_0 = \mathbf{R}(\mathbf{I} - \mathbf{A}_1) - \mathbf{R}^2 \mathbf{A}_2$. Substituting this into (16), we obtain

$$\begin{aligned}
\mathbf{D}(z) &= -z^2 \mathbf{R}(\mathbf{A}_1 - \mathbf{I}) - z^2 \mathbf{R}^2 \mathbf{A}_2 + z(\mathbf{A}_1 - \mathbf{I}) + \mathbf{A}_2 \\
&= (\mathbf{I} - z\mathbf{R})(z(\mathbf{A}_1 - \mathbf{I}) + (\mathbf{I} + z\mathbf{R})\mathbf{A}_2) \\
&= (\mathbf{I} - z\mathbf{R})(z(\mathbf{U} - \mathbf{I}) + \mathbf{A}_2) \\
&= (\mathbf{I} - z\mathbf{R})(\mathbf{U} - \mathbf{I})\left(z\mathbf{I} - (\mathbf{I} - \mathbf{U})^{-1}\mathbf{A}_2\right).
\end{aligned}$$

The third equality follows from (25), with $\mathbf{U} = \mathbf{A}_1 + \mathbf{R}\mathbf{A}_2$. Because the determinant of a matrix product is the product of the individual determinants, we gather that the roots of $\det \mathbf{D}(z)$ are simply given by the roots of its factors. The stable modes of $\mathbf{\Pi}(z)$ correspond to the roots of $\det(\mathbf{I} - z\mathbf{R})$. Since $(\mathbf{I} - \mathbf{U})^{-1}\mathbf{A}_2$ is a stochastic matrix, the Perron–Frobenius theorem asserts that

$$\det\left(z\mathbf{I} - (\mathbf{I} - \mathbf{U})^{-1}\mathbf{A}_2\right)$$

has a root at unity, with any other root having a magnitude smaller than 1. These remaining roots correspond to unstable modes of $\mathbf{\Pi}(z)$. Under partial fraction expansion, the stability constraint forces the coefficients associated with these latter roots to vanish. The remaining unknown is resolved through the normalization axiom of probability laws. This reconciles the two approaches, which necessarily lead to the same solution. The generating function method can give closed-form expressions if the channel has only two states, whereas the matrix-geometric method gives rise to a numerical procedure that works well for any finite-state channel.

V. Performance Evaluation

The detailed characterization presented in the previous sections makes it possible to compute a number of performance criteria for the system under consideration, including the probability of decoding failure, average throughput, and mean delay. In this paper, we focus on two additional performance measures relevant to delay-sensitive communications. We consider the probability that the queue occupancy exceeds a certain threshold, $\Pr(Q > \tau)$. Furthermore, we examine the decay rate of the complementary cumulative distribution function of the queue, as presented in Corollary 4.4.

Throughout this numerical study, unless stated otherwise, we employ the following system parameters. The Gilbert–Elliott erasure channel is defined by $\alpha = 0.02$, $\beta = 0.005$, $\varepsilon_1 = 0.49$, and $\varepsilon_2 = 0.0025$. This yields an average bit-erasure probability of $\overline{\varepsilon} = 0.1$ and the channel memory decays at an exponential rate of $(1 - \alpha - \beta) = 0.975$. During every codeword transmission attempt, a new packet arrives at the source with probability $\gamma = 0.25$, and the expected packet length is set to $\rho^{-1} = 195$ bits. The block length is fixed at $N = 114$ symbols. If codewords are transmitted every 4.615 ms, then this corresponds to a mean arrival rate of roughly 10.6 kb/s and an ergodic channel capacity of roughly 22.2 kb/s. These quantities are selected to loosely reflect the operation of a wireless GSM-based relay link. Collectively, these parameters dictate the evolution of the Markov process governing the queue.

The Shannon capacity for the Gilbert–Elliott erasure channel is $1 - \overline{\varepsilon}$ bits per channel use. This limit can be achieved using a sequence of independent and uniformly distributed random variables. In fact, this statement remains true for an arbitrary (ergodic) binary erasure channel where the expected number of erasures is independent of the input sequence. Suppose X and Z denote the input and output vectors of an erasure channel, respectively. Let O be a vector that indicates the observed (i.e., not erased) positions at the destination. Then, we can write

$$\begin{aligned}
I(X;Z) &= H(Z) - H(Z|X) \\
&= H(O) + H(X_O|O) - H(O) \\
&= H(X_O|O)
\end{aligned}$$

where X_O is the subvector of X that contains the value of every observed symbol. The second equality follows from the fact that there is a natural bijection between the set of possible outcome vectors and admissible pairs of the form (O, X_O). The conditional entropy $H(X_O|O = o)$ is uniformly maximized by drawing input X from a uniform distribution. Consequently, the mutual information is also maximized by choosing X according to a uniform distribution. Hence, the maximum mutual-information rate is equal to the average number of unerased positions.

We continue our analysis with simple performance criteria that are based solely on the evolution of the channel. They do

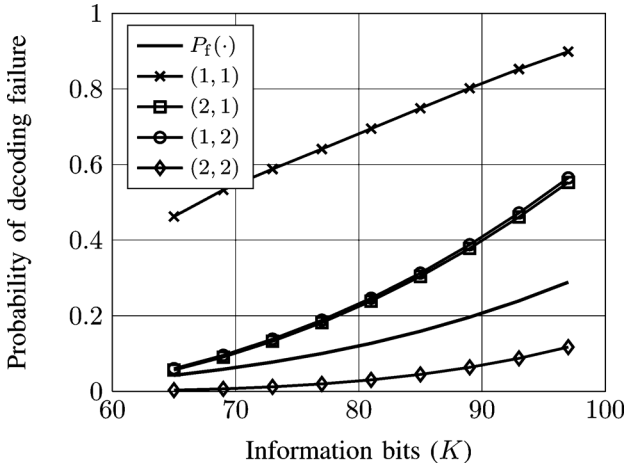

Fig. 3. Probability of decoding failure, $P_{\mathrm{f}}(N-K)$, as a function of the number of information bits per codeword. The conditional probabilities of decoding failure for various values of (c_1, c_{N+1}) are also included.

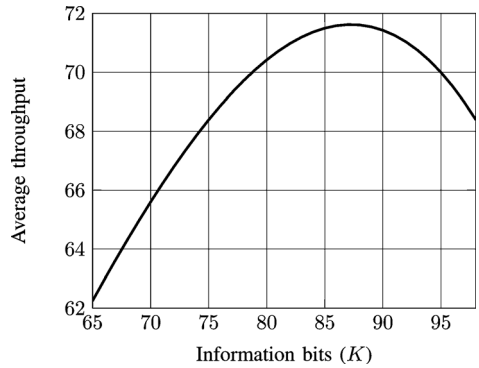

Fig. 4. Average throughput for a saturated source as a function of K, the number of information bits per codeword. The maximum throughput is obtained at $K = 87$.

not take into consideration the behavior of the queue at the transmitter. One such criterion is the probability of decoding failure at the receiver, which is equal to

$$P_{\mathrm{f}}(N-K) = \sum_{e=0}^{N} P_{\mathrm{f}}(N-K,e)\Pr(E=e)$$
$$= \sum_{e=0}^{N} P_{\mathrm{f}}(N-K,e)[\![x^e]\!]\left(\begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} \mathbf{P}_x^N \mathbf{1}\right).$$

A closely related measure of performance is the average throughput associated with a saturated source. In the present setting, this is given by

$$T_s(K,N) = \frac{K}{N}\left(1 - P_{\mathrm{f}}(N-K)\right).$$

The probability of decoding failure as a function of K appears in Fig. 3. The average throughput associated with a saturated source is plotted as a function of the number of information bits per codeword in Fig. 4. These two figures illustrate well the tradeoff between data content and error protection. In particular, these competing considerations lead to the unimodal throughput function of Fig. 4, where optimal performance is achieved at $K = 87$. A naïve conjecture would place $K = rN$ close to the rate implied by the Shannon limit $(1-\overline{\varepsilon})N = 0.9 \times 114 = 102.6$, but this is much larger than the throughput-maximizing value of $K = 87$. This observation reinforces the claim that insights gained from information theory must be modified for communication systems subject to stringent delay restrictions.

When using a short block length, two factors affect the optimal code rate K for a prescribed queueing performance. The block length may be too small to ensure convergence of the empirical average number of erasures within a block. In addition, dependencies from block to block may not be negligible. Although the probability of decoding failure and the average throughput account for channel correlation within a block, they do not capture dependencies from block to block. This is a subtle yet important observation, especially for delay-sensitive traffic. The impact of these factors becomes more severe with increasing channel memory. This consideration underlies much of the queueing analysis presented in this paper. Time dependencies in the service process of a queue can alter system performance dramatically. We, thus, turn to queue-based performance criteria.

Fig. 5 depicts the complementary cumulative distribution function of the queue, $\Pr(Q > \tau)$, as a function of K. Each curve represents the probability that, in steady state, buffer occupancy exceeds a certain threshold τ, where $\tau \in \{5, 10, 15, 20, 25\}$. The low threshold values reflect the intended use of this methodology in the context of delay-sensitive applications. As expected, the probability of the queue being greater than a prescribed threshold decreases as τ increases. More interestingly, we note that $K = 83$ appears uniformly optimal for all values of τ. That is, the best code rate seems impervious to the choice of threshold value τ. This robustness property remains present for the other system parameters we tested. Further supporting evidence for this observation is offered by looking at the asymptotic decay rate in tail occupancy, displayed in Fig. 6. When the arrival rate $\gamma\rho^{-1}$ is between 47.5 and 60, one finds that $K = 83$ is also optimal in terms of tail decay. The true optimum $K = 83$ is closer to the throughput maximizing code rate $K = 87$ than to the naïve conjecture.

We explore the impact of channel correlation on optimal code rate in our next set of results. We fix $\beta : \alpha$ at a ratio of one to four, and vary the memory factor $(1-\alpha-\beta)$. When the channel is memoryless, the optimal K is 93. For comparison, the capacity is $1-\overline{\varepsilon} = 0.9$, which yields K of roughly 103. As correlation increases, the optimal value of K initially decreases, thereby offering more protection against erasures. Yet, when the coherence time of the channel starts to approach the length of a codeword, $N = 114$, the error-correcting code becomes ineffective as it fails to handle the increasingly likely long sequences of successive erasures. The optimal strategy then progressively shifts to including more information bits in every packet, and hoping that the channel remains in its good state. In the limiting regime where $(1-\alpha-\beta)$ approaches 1, the optimal strategy is to transmit uncoded data, i.e., $K = N$. Indeed, strong correlation is characterized by long strings of erasures followed by longer strings of reliable bits, and the best strategy is to send as many information bits as possible when the channel is good. At

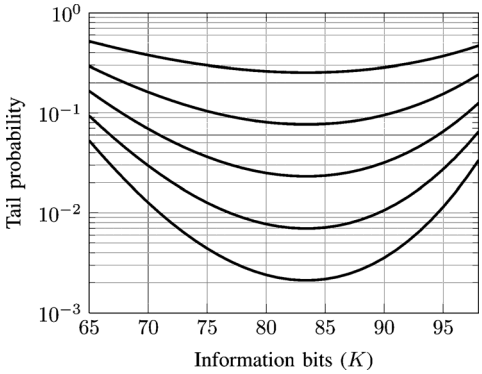

Fig. 5. Tail probabilities in the equilibrium packet distribution of the queue $\Pr(Q > \tau)$, for threshold values $\tau \in \{5, 10, 15, 20, 25\}$, as functions of the number of information bits K per codeword. The minimums occur uniformly at $K = 83$ for all threshold values.

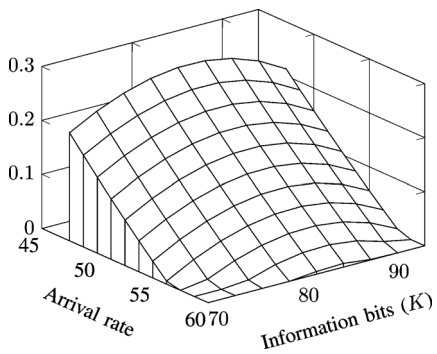

Fig. 6. Tail decay rate, $-\lim_{\tau\to\infty} \frac{1}{\tau} \log \Pr(Q > \tau)$, as a function of the number of information bits K per codeword and the average arrival rate $\gamma\rho^{-1}$ in bits per codeword transmission interval.

TABLE I
OPTIMAL NUMBER OF INFORMATION BITS PER CODEWORD AND THRESHOLD VIOLATION PROBABILITY AS FUNCTIONS OF THE CHANNEL MEMORY PARAMETER

$1-\alpha-\beta$	Optimal K	$\min_K \Pr(Q > 5)$
0	93	0.0359
0.5	91	0.0438
0.9	85	0.0982
0.98	85	0.2843
0.99	95	0.3169

this point, the bit erasure channel essentially becomes a correlated packet erasure channel. Numerical results are summarized in Table I.

VI. DISCUSSION AND CONCLUDING REMARKS

This paper presents a new framework to analyze the relation between code rate and queueing behavior for communications over channels with memory. The simplicity of the erasure channel and its closed-form characterization of error events are instrumental in conducting our analysis. For short block length and channels with memory, the optimal code rate appears to be linked to the relative size of a codeword compared to the coherence time of the channel. In certain circumstances, it is beneficial to provide significant protection against erasures. However, as channel memory increases, performance may be improved by incorporating more data bits in every codeword. In this latter case, the transmitter resorts to a strategy where information is successfully sent when the channel starts in a good state, and it is lost otherwise. This is in stark contrast to information-theoretic results obtained through asymptotically long codewords. Once the block length is selected, the optimal code rate seems rather insensitive to the queue occupancy threshold. This observation considerably simplifies system design because an optimal code rate can be selected irrespective of the target queue length. The set of admissible arrival rates, on the other hand, will depend heavily on the queueing objective.

A distinguishing feature of this work is that it provides a rigorous approach linking queueing performance to the operation of a communication system at the physical layer. The methodology and results are developed for the Gilbert–Elliott erasure channel, but can be generalized to more intricate finite-state channels with memory. For example, the simple performance characterization of random codes over erasure channels may extend to hard-decision decoding of BCH codes over Gilbert–Elliott error channels. Possible avenues of future research include the study of alternative arrival processes, the ability to vary the rate and the length of codewords dynamically, and exploring pragmatic feedback schemes.

APPENDIX

This appendix contains demonstrations for Theorem 4.2 and Corollary 4.4.

A. Proof of Theorem 4.2

For completeness, we provide a succinct outline of a proof to this theorem; our arguments are motivated, partly, by the derivation presented by Latouche and Ramaswami [34]. For quasi-birth-death processes, several authors have reported similar results [27], [37], [29].

The transitions of the Markov chain $\{Y_s\}$, excluding states at level zero, are governed by the substochastic matrix

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{26}$$

We note that the transitions of levels in $\{Y_s\}$, excluding levels 0 and 1, are dictated by the same semi-infinite matrix (26). Exploiting this symmetry and the fact that $\{Y_s\}$ can only jump to neighboring levels, one can use the definition of $\mathbf{U}$ to obtain the following implicit equation:

$$\begin{aligned} \mathbf{U} &= \mathbf{A}_1 + \mathbf{A}_0 \left(\sum_{i=0}^{\infty} \mathbf{U}^i \right) \mathbf{A}_2 \\ &= \mathbf{A}_1 + \mathbf{A}_0 \left(\mathbf{I} - \mathbf{U} \right)^{-1} \mathbf{A}_2. \end{aligned} \tag{27}$$

This is equivalent to a quadratic matrix equation and it can be solved efficiently using numerical methods. For instance, multiplying both sides of (27) by $\mathbf{R}$, substituting $\mathbf{R} = \mathbf{A}_0 \left(\mathbf{I} - \mathbf{U} \right)^{-1}$, and rearranging terms, we obtain

$$\mathbf{A}_0 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2\mathbf{A}_2 = \mathbf{R}. \tag{28}$$

We proceed to show that the iterative algorithm of (22) is one possible method to obtain $\mathbf{U}$. Consider the probability space of paths on $\{1,2\}\times\mathbb{N}_0$ with the measure induced by Markov chain $\{Y_s\}$. Let $S_j(c,d)$ be the event such that, starting from state $(1,c)$, the Markov chain $\{Y_s\}$ first re-enters level 1 by visiting $(1,d)$ and with the excursions constrained to lie between levels 1 and j. It follows from this definition that $S_j(c,d)\subset S_{j+1}(c,d)$ and, therefore

$$\lim_{j\to\infty} S_j(c,d) = \cup_{j\in\mathbb{N}} S_j(c,d).$$

Utilizing the monotone convergence theorem, we gather that

$$\lim_{j\to\infty} \Pr(S_j(c,d)) = \Pr\left(\bigcup_{j\in\mathbb{N}} S_j(c,d)\right).$$

By construction, $[\mathbf{U}]_{c,d} = \Pr\left(\bigcup_{j\in\mathbb{N}} S_j(c,d)\right)$. To complete the proof, it remains to show that $[\mathbf{U}_j]_{c,d}$, as defined in (22), is the probability of event $S_j(c,d)$.

Formally, this is equivalent to the mathematical statement $[\mathbf{U}_j]_{c,d} = \Pr(S_j(c,d))$ for all $j\in\mathbb{N}$, which we verify using induction. From the definition of $S_j(c,d)$, we immediately obtain $\Pr(S_1(c,d)) = [\mathbf{A}_1]_{c,d}$ and, consequently, $[\mathbf{U}_1]_{c,d} = \Pr(S_1(c,d))$ because $\mathbf{U}_1 = \mathbf{A}_1$. To continue, we assume this proposition holds for all integers less than or equal to j and show this implies that it holds for $j+1$. First, we note that $S_{j+1}(c,d)$ is the event such that, starting from state $(1,c)$, the Markov chain $\{Y_s\}$ first re-enters level 1 by visiting $(1,d)$ and with the excursions constrained to lie between levels 1 and $j+1$. The elements in $S_{j+1}(c,d)$ can be partitioned into sets according to their number of visits to level 2. In particular, the Markov chain $\{Y_s\}$ remains at level 1 with probability $[\mathbf{A}_1]_{c,d}$. Alternatively, it can immediately transition to level 2, revisit this level a number of times while remaining between levels 2 and $j+1$, and then jump back down to level 1.

Key to the proof is the symmetric nature of the chain: the probability that, starting from state $(2,c)$, the Markov chain $\{Y_s\}$ first re-enters level 2 by visiting $(2,d)$ and with the excursions constrained to lie between levels 2 and $j+1$ is equal to $\Pr(S_j(c,d))$. Indeed, there is a natural, probability-preserving bijection between paths in $S_j(c,d)$ and paths from $(2,c)$ that first re-enter level 2 at $(2,d)$ and remain between levels 2 and $j+1$. By the Markov property and our inductive hypothesis, we can write the probability that, starting from $(1,c)$, the Markov chain immediately goes up to level 2, and visits this level exactly $k+1$ times before it first re-enters level 1 at $(1,d)$ as $\left[\mathbf{A}_0\mathbf{U}_j^k\mathbf{A}_2\right]_{c,d}$. The proposed partition of $S_{j+1}(c,d)$ is a countable union of disjoint events, where each set accounts for a distinct number of visits to level 2. It follows from the renewal property of Markov chains and the symmetry of the problem that

$$\begin{aligned}\Pr\left(S_{j+1}(c,d)\right) &= \left[\mathbf{A}_1 + \mathbf{A}_0\sum_{k=0}^{\infty}\mathbf{U}_j^k\mathbf{A}_2\right]_{c,d}\\ &= \left[\mathbf{A}_1 + \mathbf{A}_0\left(\mathbf{I}-\mathbf{U}_j\right)^{-1}\mathbf{A}_2\right]_{c,d} = [\mathbf{U}_{j+1}]_{c,d}\end{aligned}$$

where the second equality follows from the Neumann expansion and the third equality is an application of definition (22). Hence, for every $j\in\mathbb{N}$, we have $\Pr(S_j(c,d)) = [\mathbf{U}_j]_{c,d}$. This establishes that the iterative algorithm of (22) converges to $\mathbf{U}$, as desired.

To complete the proof, it remains to show that the candidate distribution specified in Theorem 4.2 is indeed the invariant distribution of $\mathbf{T}$. Notice that (28) immediately ensures that

$$\pi_{q-1}\mathbf{A}_0 + \pi_q\mathbf{A}_1 + \pi_{q+1}\mathbf{A}_2 = \pi_q \tag{29}$$

for $q\geq 2$. Consider the finite matrix $\tilde{\mathbf{T}}$ with nonnegative entries introduced in (23). We wish to prove that this is a stochastic matrix. Since $\mathbf{T}$ represents a probability transition matrix, we already have $(\mathbf{C}_0+\mathbf{C}_1)\mathbf{1}=\mathbf{1}$. To establish the second equality, we examine the following progression:

$$\begin{aligned}(\mathbf{A}_2+\mathbf{A}_1+\mathbf{R}\mathbf{A}_2-\mathbf{I})\mathbf{1} &= (\mathbf{R}\mathbf{A}_2-\mathbf{A}_0)\mathbf{1}\\ &= \left(\mathbf{R}\mathbf{A}_2+\mathbf{R}\mathbf{A}_1+\mathbf{R}^2\mathbf{A}_2-\mathbf{R}\right)\mathbf{1}\\ &= \mathbf{R}(\mathbf{A}_2+\mathbf{A}_1+\mathbf{R}\mathbf{A}_2-\mathbf{I})\mathbf{1}\end{aligned}$$

where the first step relies on the identity $(\mathbf{A}_0+\mathbf{A}_1+\mathbf{A}_2)\mathbf{1}=\mathbf{1}$ and the second step follows from (28). Since the matrix $\mathbf{I}-\mathbf{R}$ is invertible, one can move all terms to the left-hand side to see that $(\mathbf{A}_2+\mathbf{A}_1+\mathbf{R}\mathbf{A}_2-\mathbf{I})\mathbf{1}=\mathbf{0}$ and therefore $(\mathbf{A}_2+\mathbf{A}_1+\mathbf{R}\mathbf{A}_2)\mathbf{1}=\mathbf{1}$. That is, (23) is a stochastic matrix. This implies that it admits an invariant distribution, which must satisfy $\tilde{\pi}_0\mathbf{C}_1+\tilde{\pi}_1\mathbf{A}_2=\tilde{\pi}_0$ and $\tilde{\pi}_0\mathbf{C}_0+\tilde{\pi}_1(\mathbf{A}_1+\mathbf{R}\mathbf{A}_2)=\tilde{\pi}_1$. Then, for the distribution defined in (24), we get

$$\begin{aligned}\pi_0\mathbf{C}_1+\pi_1\mathbf{A}_2 &= \pi_0\\ \pi_0\mathbf{C}_0+\pi_1\mathbf{A}_1+\pi_2\mathbf{A}_2 &= \pi_0\mathbf{C}_0+\pi_1(\mathbf{A}_1+\mathbf{R}\mathbf{A}_2)=\pi_1.\end{aligned}$$

These equations, together with (29), imply that the distribution defined in (24) is invariant under $\mathbf{T}$, as desired.

B. Proof of Corollary 4.4

Since $\mathbf{R}$ is a positive matrix, the Perron–Frobenius theorem implies that $\mathbf{R}$ has a unique positive eigenvalue $\lambda=\varrho(\mathbf{R})$ of maximum modulus [32]. Furthermore, this eigenvalue is associated with a positive left eigenvector v, and a positive right eigenvector w. It follows that

$$\frac{u\mathbf{R}^j}{\varrho(\mathbf{R})^j} = u\left(\frac{w^Tv}{vw^T}+o(1)\right) = \frac{uw^T}{vw^T}\left(v+o(1)\right)$$

for any nonnegative, nonzero vector u. For any integer τ, the tail probability of the queue is consequently governed by

$$\begin{aligned}\Pr(Q>\tau) &= \pi_1\left(\sum_{q=\tau}^{\infty}\mathbf{R}^q\right)\mathbf{1} = \pi_1\mathbf{R}^\tau(\mathbf{I}-\mathbf{R})^{-1}\mathbf{1}\\ &= \frac{\pi_1w^T}{vw^T}\varrho(\mathbf{R})^\tau\left(\left(v+o(1)\right)(\mathbf{I}-\mathbf{R})^{-1}\mathbf{1}\right).\end{aligned}$$

Taking the normalized limit of the logarithm completes the proof.

References

[1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ: Wiley, 1968.

[3] D. J. Costello Jr., J. Hagenauer, H. Imai, and S. B. Wicker, "Applications of error-control coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2531–2560, Oct. 1998.

[4] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.
[5] R. A. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.
[6] S. Shakkottai, "Effective capacity and QoS for wireless scheduling," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 749–761, Apr. 2008.
[7] L. Ying, S. Yang, and R. Srikant, "Optimal delay-throughput tradeoffs in mobile ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4119–4143, Sep. 2008.
[8] M. V. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Trans.*, vol. 12, no. 4, pp. 250–265, 1976.
[9] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 25, no. 6, pp. 729–733, Nov. 1979.
[10] P. Berlin, B. Nakiboglu, B. Rimoldi, and E. Telatar, "A simple converse of Burnashev's reliability function," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3074–3080, Jul. 2009.
[11] 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.
[12] S. Goel and R. Negi, "Analysis of delay statistics for the queued-code," in *Proc. IEEE Int. Conf. Commun.*, Dresden, Germany, Jun. 2009, pp. 1–6.
[13] Y.-C. Ko, M.-S. Alouini, and M. K. Simon, "Outage probability of diversity systems over generalized fading channels," *IEEE Trans. Commun.*, vol. 48, no. 11, pp. 1783–1787, Nov. 2000.
[14] L. Li, N. Jindal, and A. Goldsmith, "Outage capacities and optimal power allocation for fading multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1326–1347, Apr. 2005.
[15] J. G. Kim and M. M. Krunz, "Bandwidth allocation in wireless networks with guaranteed packet-loss performance," *IEEE/ACM Trans. Netw.*, vol. 8, no. 3, pp. 337–349, Jun. 2000.
[16] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.
[17] 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.
[18] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the PHY be?," in *Proc. IEEE Global Telecommun. Conf.*, Honolulu, HI, Dec. 2009, pp. 1–6.
[19] 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.
[20] 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.
[21] 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.
[22] P. Parag, J.-F. Chamberland, H. D. Pfister, and K. R. Narayanan, "Code rate, queueing behavior and the correlated erasure channel," in *Proc. IEEE Inf. Theory Workshop*, Cairo, Egypt, Jan. 2010, pp. 1–6.
[23] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[24] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. New York: Wiley-Interscience, 1975.
[25] F. Baccelli and P. Bremaud, *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, 2nd ed. New York: Springer-Verlag, 2003.
[26] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, ser. Probability and Statistics, 4th ed. New York: Wiley-Interscience, 2008.
[27] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, ser. Probability: Pure and Applied. Boca Raton, FL: CRC Press, 1989.
[28] G. Latouche and V. Ramaswami, "A logarithmic reduction algorithm for quasi-birth-death processes," *J. Appl. Probab.*, vol. 30, no. 3, pp. 650–674, Sep. 1993.
[29] B. Hajek, "Birth-and-death processes on the integers with phases and general boundaries," *J. Appl. Probab.*, vol. 19, no. 3, pp. 488–499, Sep. 1982.
[30] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, ser. Graduate Texts in Mathematics, 2nd ed. New York: Springer-Verlag, 1976.
[31] J. R. Norris, *Markov Chains*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[32] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[33] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. New York: Dover, 1995.
[34] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ser. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia, PA: SIAM, 1987.
[35] D. M. Lucantoni, "The BMAP/G/1 queue: A tutorial," in *Performance Evaluation of Computer and Communications Systems*, ser. Lecture Notes in Computer Science, L. Donatiello and R. Nelson, Eds. New York: Springer Verlag, 1993, pp. 330–358.
[36] A. S. Alfa and W. Li, "Matrix-geometric analysis of the discrete time GI/G/1 system," *Stoch. Models*, vol. 17, no. 4, pp. 541–554, 2001.
[37] R. V. Evans, "Geometric distribution in some two-dimensional queuing systems," *Oper. Res.*, vol. 15, no. 5, pp. 830–846, Sep.–Oct. 1967.
[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley-Interscience, 2006.
[39] 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.

Parimal Parag (S'05) received the dual degree (B.Tech. and M.Tech.) in electrical engineering, with specialization in communication systems, from the Indian Institute of Technology at Madras in 2004. He received the Ph.D. degree in 2011 from Texas A&M University, College Station. He was a visiting scholar at the Los Alamos National Laboratory in 2007, and at Stanford University in 2010. His research interests include applied probability, statistical signal processing, queueing theory, information theory, optimization methods, estimation and detection theory, and their application to communication networks. He was a silver medalist at the Indian Institute of Technology, Madras. He is also a recipient of Indian National Talent Search Scholarship.

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 the Ph.D. degree in electrical engineering from the University of California at San Diego in 2003 and joined the faculty of the College of Engineering at Texas A&M University in 2006. Prior to that, he spent two years in research and development at Qualcomm, Inc. and one year as a postdoctoral fellow at the Ecole Polytechnique Fédérale de Lausanne (EPFL). He received a CAREER Award from the National Science Foundation 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.

Krishna Narayanan (S'92–M'98–SM'09) received the Ph.D. degree in electrical engineering from the Georgia Institute of Technology in 1998 and is currently a Professor in the Department of Electrical and Computer Engineering at Texas A&M University. His research interests are in coding theory, information theory, joint source-channel coding and signal processing with applications to wireless communications and data storage. He was the recipient of the 2006 Best Paper Award from the IEEE Technical Committee for Signal Processing for Data Storage. He served as the area editor for the coding theory and applications area of the IEEE TRANSACTIONS ON COMMUNICATIONS from 2007 until 2011.