Abstract—Data transmission over a discrete memoryless channel is considered when the arrival of data is bursty and is subject to a delay deadline. An exponential decay of the probability of delay violation with respect to a large delay deadline is proved when the block length scales linearly with the deadline. When considered in conjunction with Gallager's error exponents, the first natural consequence of this result is a separation principle: a separated scheme of buffering traffic and block-coding transmissions achieves arbitrarily high reliability for an asymptotically large delay budget. Furthermore, the exponential decay nature of the result provides insight as how to budget the delay limit between the coding time and the waiting time in the queue.

I. INTRODUCTION

Shannon’s classical information theory [1] states that it is feasible to communicate information over a noisy channel with arbitrarily high reliability at all rates below channel capacity. Gallager’s error exponents [3] quantify this reliability asymptotically. However, if the information is arriving at the transmitter in a stochastic manner, Shannon’s coding schemes will make it necessary for the transmitter to wait arbitrarily long prior to transmission. This is not a palatable solution to the problem of information delivery in delay sensitive applications such as voice over IP and video streaming. In fact, in many applications of interest, the utility of information vanishes beyond an acceptable deadline known as the delay budget. In such applications one can identify two sources of information loss: the first is the erroneous decoding at the receiver as a result of channel errors, and the second one is the tardy decoding of bits past the deadline. In a conventional communication system with block coding and the absence of feedback, these sources of loss are to be determined and traded off via an optimized choice of block coding length. Intuitively, a longer block length makes channel induced errors less probable, whereas it increases the probability of tardy decoding of data. In this work, we study this tradeoff asymptotically i.e., we ask, “given a large delay deadline, what should the block length be so that the overall loss probability is minimized?”.

To answer the above question concretely, one requires a large deviation analysis of a single-server queue with bulk service when the bulk size scales to infinity. Such an analysis do not arise in traditional queuing systems [2]. The present work, to the best of our knowledge, for the first time provides a large deviation principle for the delay in a single-server queue with bulk service. An important consequence of our result when combined with Gallager’s error exponents is a separation of queuing and coding principle. Furthermore, our analysis provides insight as how to optimally budget the delay limit between coding duration and waiting time in the queue.

Motivated by cross-layer design of networks, our work belongs to a growing literature on the joint analysis and design of queuing and communication schemes [4]-[9]. In particular, our work compliments [5]-[9] where tradeoffs between delay, power and throughout are considered in the context of fading channels. Our work also compliments the results of [4] where a LDP analysis of a single server bulk service queue is presented but for a finite bulk size and asymptotically many flows.

This paper is organized as follows: Section II describes the system model. Section III proves the exponential decay of probability of delay violation with linear scaling of the block length. Section IV talks about the combined error exponent of block error and delay violation. Finally, Section V concludes the paper.

The notations used in the paper are as follows: $\mathbb{R}$, $\mathbb{R}^+$ and $\mathbb{Z}$ denote the set of real numbers, positive real numbers and integers respectively. Also, $\mathbb{R}_+^* = \{\mathbb{R}_+ \cup \infty\}$

II. SYSTEM MODEL

We consider a slotted communication system depicted in Fig 1. Bits arrive at the transmitter in a stochastic and bursty manner with average rate $\lambda$ and are to be transmitted over a discrete memoryless channel (DMC) with capacity $C > \lambda$. The bits are queued up in an infinite buffer and are served in a first-come first-serve discipline. In addition, bits are coded in blocks of length $N$ at rate $\lambda < R < C$ prior to transmission. Consequently, every $N$ time-slots, the $RN$ oldest bits are instantaneously removed from the queue and are transmitted
over the next $N$ time-slots.\(^1\) The delay for a bit arriving at the transmitter in time slot $t_1$ and being decoded at the receiver in time slot $t_2$ is defined to be $t_2 - t_1$ slots.

**Failure event:** A bit is said to have failed transmission if 1) it is in a block which was erroneously decoded or 2) the delay for the bit exceeds a parameter $D$ (delay budget). Since $\lambda < R < C$, it is meaningful to speak of a steady state operation, in which the average probability of delay violation is denoted by $P_{\text{delay}}$, probability of a decoding error is denoted by $P_{\text{code}}$ and the probability of a failure event is denoted by $P_{\text{failure}}$. Note that $P_{\text{failure}}$ represents the reliability of the communication system in presence of realistic quality of service requirements.

**Problem statement:** Given the above communication system, we are interested in characterizing $P_{\text{failure}}$ as a function of the delay budget $D$, coding length $N$ and coding rate $R$. For the sake of analytic tractability, we consider an asymptotic analysis in which $D, N \to \infty$.

**Scaling of the block length:** In this work, we restrict our attention to linear scalings of the block length with the delay budget, i.e. we assume $N = \lceil \frac{D}{b} \rceil$, where $b > 2$ is a dimensionless parameter of interest.

The arrival stochastic process is denoted as $\{A_t\}_{t \in \mathbb{Z}}$, where $A_t$ represents the number of bits arriving in the time interval $[t, t + 1)$. Let $S_n = \sum_{t=1}^{n} A_t$ and $\Lambda_n(\theta) = \log E(e^{\theta S_n})$. We assume that the arrival process is stationary and obeys a large deviation principle. More precisely, we assume that the arrival process is such that it satisfies the Gartner-Ellis Theorem:

**Assumption 1.** Arrival process $A_t$ is stationary and its limiting scaled log normal moment generating function satisfies:

$$\lim_{n \to \infty} \frac{1}{n} \log E(e^{\theta S_n}) = \lim_{n \to \infty} \frac{1}{n} \Lambda_n(\theta) = \Lambda(\theta). \quad (1)$$

Generating function $\Lambda(\theta) \in \mathbb{R}_+ \forall \theta \in \mathbb{R}_+$ is called the cumulant generating function. We also utilize its convex conjugate, $\Lambda^*(x)$, defined as

$$\Lambda^*(x) = \sup_{\theta \in \mathbb{R}} \theta x - \Lambda(\theta) \quad (2)$$

**Remark 1:** Assumption 1 is a relatively weak condition on the arrival process and is satisfied for a wide array of traffic systems such as those with iid increments, Markov modulated arrivals, compound Poisson process [2].

The contributions of our work are:

1) We prove that the probability of delay violation $P_{\text{delay}}(D, \Lambda, R, b)$ decays exponentially with $D$ and find the appropriate rate function.

2) Given fixed parameters $\Lambda, R$ and $b$, we prove that the probability of failure $P_{\text{failure}}(D, \Lambda, R, b)$ decays exponentially with $D$. This establishes a queueing-coding separation principle.

3) We find the optimal scaling parameter $b^*$ for the best $P_{\text{failure}}$ performance, hence giving insights into an important question as to how (asymptotically and in a scaling sense) budget delay between the queue and the channel

**III. LARGE DEVIATION ANALYSIS OF BULK QUEUES**

In this section, we present a large deviation analysis of bulk service queues with the bulk size scaling linearly with the delay limit. This large deviation analysis, providing an asymptotic expression for $P_{\text{delay}}$ in a communications system with block coding, is the main technical contribution of our work and, to the best of our knowledge, is the first such study in the field of queuing theory and communication theory.

The results of this section can be summarized in the theorem below:

**Theorem 1.** Consider the communication system described in Section II with input arrival process $A_t$, satisfying Assumption 1, cumulant generating function $\Lambda(\theta)$, fixed and known coding rate $R$, delay budget $D$, coding block length $N = \lceil \frac{D}{b} \rceil$ and a fixed scaling parameter $b > 2$. Then the delay event exhibits a large deviation, i.e.:

$$\lim_{D \to \infty} \frac{\log \Pr(\text{delay} > D)}{D} = W(R, b) \quad (3)$$

where:

$$W(R, b) = -R \frac{[b - 1]}{b} \left(1 - \frac{1 - b + [b]}{[b - 1]}\right) I(1) \quad (4)$$

and

$$I(q) = \inf_{t \in \mathbb{R}_+} t \Lambda^*(R + q/t) \quad (5)$$

**Remark 2:** As one would expect, at large values of $b$, the effect of the bulk service is negligible and (3) reduces to the rate function of a single server queuing system without batching [2]. However, as $b$ is brought closer to 2, the effect of bulk service starts to set in and significantly reduces the exponent. Note that when the coding time is greater than $\frac{D}{b}$, delay violation will occur with a probability of at least $\frac{2}{b}$ (and hence not exponential) as bits arriving in the timeslot right after service will have a delay violation with probability 1; hence our requirement that $b > 2$.

**Proof of Theorem 1:**

Along any sequence $D \in \mathbb{R}$ that goes to $\infty$, $bN \leq D \leq bN + b$. This implies that:

$$\liminf_{D \to \infty} \frac{\log \Pr(\text{delay} > D)}{D} \geq \liminf_{N \to \infty} \frac{\log \Pr(\text{delay} > bN + b)}{bN} \quad (6)$$

and

$$\limsup_{D \to \infty} \frac{\log \Pr(\text{delay} > D)}{D} \leq \limsup_{N \to \infty} \frac{\log \Pr(\text{delay} > bN)}{bN + b} \quad (7)$$

Next, by means of Lemma 1, we show that the right hand sides of both equations (6) and (7) are equal to $W(R, b)$.

**Lemma 1.** In the setting of theorem 1, the following result
In the setting of theorem 1 and for all $i$ time slot, consequently, we focus on the last bit that arrives in time slot $i$ probability of delay for the last bit that arrives in time slot $i$. Moreover, it is easy to show that in our setting and our scale of interest, the average probability of delay for bits that arrive in time slot $i$ is equal to the probability of delay for the last bit that arrives in that time slot. Consequently, we focus on the last bit that arrives in time slot $i$ (see Sec III.A of [4]); denote the delay encountered by bits arriving at time slot $i$ by $\phi$.

Since the input arrival process $A_t$ is stationary, it is equally likely that a bit arrives in any of the $N$ slot types. In other words, the probability of delay violation for any fixed value $\phi$ and $b$ given by:

$$\Pr(\text{delay} > bN + \phi) = \frac{1}{N}\sum_{i=1}^{N} \Pr(\text{delay}_i > bN + \phi)$$

(9)

For a given $N$, define:

$$i_N^* = \arg \max_{1 \leq i \leq N} \Pr(\text{delay}_i > bN + \phi)$$

(10)

Then it follows that:

$$\liminf_{N \to \infty} \frac{1}{N} \log \left( \frac{1}{N} \sum_{i=1}^{N} \Pr(\text{delay}_i > bN + \phi) \right) \geq \liminf_{N \to \infty} \frac{\log \left( \frac{1}{N} \Pr(\text{delay}_{i_N^*} > bN + \phi) \right)}{\log \left( \Pr(\text{delay}_{i_N^*} > bN + \phi) \right)} = \liminf_{N \to \infty} \frac{\log \left( \Pr(\text{delay}_{i_N^*} > bN + \phi) \right)}{N}$$

(11)

Similarly,

$$\limsup_{N \to \infty} \frac{\log \left( \frac{1}{N} \sum_{i=1}^{N} \Pr(\text{delay}_i > bN + \phi) \right)}{N} \leq \limsup_{N \to \infty} \frac{\log \left( \Pr(\text{delay}_{i_N^*} > bN + \phi) \right)}{N}$$

(12)

However, the delay encountered by bits arriving at time slot $i$, in steady state, can be related to the steady state queue length at the end of time slot $i$ which is denoted as $Q_i$ (again note that queue lengths at time $m_1N + i$ and $m_2N + i$ are statistically identical). In other words, by careful observation of Fig. 2 (see proof of Lemma 2 in Sec III of [4]), we have the equivalence of the following events.

$$[\text{delay}_i > d] = [i + (N - i) + \frac{Q_i}{NR} > \frac{i - 1 + \phi}{N}].$$

(13)

Setting $d = bN + \phi$, we arrive at

$$\Pr(\text{delay}_i > bN + \phi) = \Pr(Q_i > RN[b-1+\frac{i-1+\phi}{N}]).$$

(14)

As shown in Fig 3, and in the steady state regime, for all $1 \leq i_2 \leq i_1 \leq N$ we have

$$Q_1 \leq Q_{i_2} \leq Q_{i_1} \leq Q_N \leq (Q_1 + NR) \leq (Q_{i_2} + NR) \ a.s$$

This implies that for all $N > [\phi]$

$$i_N^* = \max\{i : 1 \leq i \leq N, [b-1+\frac{i-1+\phi}{N}] = [b-1]\}$$

(15)

which is the same as saying:

$$i_N^* = \lfloor N(1-b+\lfloor b \rfloor) \rfloor - \lfloor \phi \rfloor$$

(16)

From (14) and (15), it follows that

$$\Pr(\text{delay}_{i_N^*} > bN + \phi) = \Pr(Q_{i_N^*} > RN[b-1]).$$

(17)

Next, we use the following lemma whose proof is given in the appendix.

**Lemma 2.** In the setting of theorem 1 and for all $\beta > R$, $a \in (0,1]$ and $k \in \mathbb{Z}$, the following limit exists and is given as:

$$\lim_{N \to \infty} \frac{1}{\beta N} \log \Pr(Q_{i_N^*} - k > \beta N) = -(1-aR)I(1).$$

(18)

Setting $\beta = R[b-1]$, $a = (1-b+\lfloor b \rfloor)$ and $k = \lfloor \phi \rfloor$ in Lemma 2 and taking the limiting case of (17), we have

$$\lim_{N \to \infty} \frac{1}{\beta N} \log \Pr(Q_{i_N^*} > bN + \phi) = -R \frac{[b-1]}{b} \left( 1 - \frac{1-b+\lfloor b \rfloor}{[b-1]} \right) I(1)$$

(19)

Inserting this back in (11) and (12) we have the assertion of the lemma.
IV. Failure Event Exponents

Let us define the following parameters which we shall refer to as exponents:

\[
E_{\text{delay}}(R, \Lambda, b) = - \lim_{D \to \infty} \frac{\log P_{\text{delay}}(D, \Lambda, R, b)}{D} \tag{20}
\]

\[
E_{\text{code}}(R) = - \lim_{N \to \infty} \frac{\log P_{\text{code}}(N, R)}{N} \tag{21}
\]

\[
E_{\text{failure}}(R, b, \Lambda) = - \lim_{D \to \infty} \frac{\log P_{\text{failure}}(D, \Lambda, R, b)}{D} \tag{22}
\]

where \( P_{\text{delay}} \), \( P_{\text{code}} \) and \( P_{\text{failure}} \) were defined in Section II.

A. Review of Gallager’s results:

Error exponents associated with block decoding against errors introduced by a discrete memoryless channel are well researched (see Chapter 5, [3]). We will briefly provide a summary of the results below:

Fact 1: For a DMC with capacity \( C \), there exist functions \( E_r : [0, C] \to \mathbb{R}_+ \), \( E_{ex} : [0, C] \to \mathbb{R}_+ \) and \( E_{sp} : [0, C] \to \mathbb{R}_+ \) such that:

\[
E_r(R) < E_{ex}(R) \leq E_{code}(R, b) \leq E_{sp}(R) \quad \forall R \in [0, C].
\]

Furthermore, there exists \( R_{cr} \in [0, C] \) such that

\[
E_r(R) = E_{sp}(R) = E_{code}(R, b) \quad \forall R \in [R_{cr}, C].
\]

Remark 3: \( E_r(R) \) is known as the random coding exponent, \( E_{ex}(R) \) is known as the expurgated exponent and \( E_{sp}(R) \) is called the sphere packing exponent.

Remark 4: If there is a non zero transition probability from every input to every output of the channel (a typical practical channel), then \( R_{cr} < C \).

B. Exponential Decay of Probability of Failure:

From Theorem 1 and (24) and (23), it is clear that \( P_{\text{delay}} \) and \( P_{\text{code}} \) decay exponentially with \( D \) when the coding block length is chosen as \( N = \lfloor \frac{D}{R} \rfloor \). This means that \( E_{\text{delay}}(R, \Lambda, b) \) and \( E_{\text{code}}(R) \) are positive for all \( R < C \) and \( b > 2 \). Furthermore, the events of decoding error and delay violation are independent. This implies:

\[
E_{\text{failure}}(R, b, \Lambda) = \min (E_{\text{delay}}(R, \Lambda, b), \frac{1}{b} E_{\text{code}}(R, b)).
\]

In other words, we have shown that \( P_{\text{failure}}(D, \Lambda, R, b) \) decays exponentially with \( D \) and in effect established the asymptotic optimality of separating queuing and coding functionalities.

C. Optimal Scaling:

In this section we look at objective 3, wherein we try to answer the question of how to budget the delay limit \( D \) between block length and queuing waiting time. Due to the nature of Gallager’s exponents, we need to consider 2 cases:

\[\text{Case 1: } R_{cr} < R < C \text{: In this case, we know the exact expressions for } E_{\text{code}}(R, b) \text{ and hence } E_{\text{failure}}(R, b, \Lambda). \text{ Because of (25) and (24), we can find } b^* \text{ by solving:}

\[
\frac{E_r(R)}{b^*} = E_{\text{delay}}(R, \Lambda, b^*)
\]

This is best demonstrated in Fig 3 which shows the solution for a binary symmetric channel supporting Poisson traffic. The solution point also gives the optimized failure exponent \( E_{\text{failure}}^*(R, \Lambda) \).

\[\text{Case 2: } 0 < R < R_{cr} : \text{ In this regime, since we do not know the exact behaviour of the decoding exponent, we cannot get exact solutions for the optimal failure exponent. Nevertheless, (23) allows us to provide upper and lower bounds. Fig 4 illustrates this graphically. By solving for an optimal } b^* \text{ for both the upper bound and lower bound of the decoding exponent, we get bounds on the optimal failure exponent } (E_{\text{failure}}^*(R, \Lambda)) \text{ as well as bounds on the optimal scaling } b^* \text{ which will have to be used. Note that for certain channels, at lower rates, } E_{sp}(R) \text{ might go to } \infty \text{ in which case the upper bound on } E_{\text{failure}}^*(R, \Lambda) \text{ will be } \infty.
\]
V. CONCLUSION

We have proved the exponential decay of probability of bit failure for a fixed rate transmission over a DMC. We have also developed ways of finding $b^*$, the optimal scaling parameter to scale the block length with the delay limit.

Interesting extensions of this work would be to explore the dependence of $E_{\text{failure}}$ on transmission rate $R$. Also of interest would be the quantification of the protocol information about the dummy bits added while coding, as discussed in [10].

**Proof of Lemma 2:** Recall that $Q_i$ denotes the steady state queue backlog (queue length) at timeslot $i = 1, 2, \ldots, N$. $Q_i$ is given by (see Lemma 1.1 of [2])

$$Q_i = \sup_{j > 0} \{SN_{j+i} - RN_j\} \quad (27)$$

Choose $N$ such that $N > k$ and let $i = [aN] - k$ where $a \in (0, 1]$. From the Chernoff bound, we have:

$$Pr(Q_{[aN] - k} > \beta N) \leq \sum_{j > 0} Pr(S_{Nj+[aN]-k} - RN_j > \beta N) \leq e^{-\beta N \theta} \sum_{j > 0} e^{\Lambda_{Nj+[aN]-k}} e^{-R N_j \theta} \quad \forall \theta > 0. \quad (28)$$

Since $E(A_t) < R$, there exists some $\theta > 0$ such that for small enough $\epsilon$,

$$\Lambda(\theta) < \theta R - 2\theta \epsilon. \quad (29)$$

In addition, (1) implies that $\exists \ell(\epsilon)$ such that whenever $(Nj + [aN] - k) > \ell(\epsilon)$

$$\Lambda_{Nj+[aN]-k}(\theta) < (Nj + [aN] - k) \Lambda(\theta) + \epsilon(Nj + [aN] - k).$$

Let $\Gamma(l, k, N) = \{j : Nj + [aN] - k > l\}$, then:

$$Pr(Q_{[aN] - k} > \beta N) \leq e^{-\beta N \theta} \sum_{j \in \Gamma(l, k, N)} e^{\Lambda_{Nj+[aN]-k}(\theta) - RN_j \theta} + e^{-\beta N \theta + [aN] \Lambda(\theta) - k \Lambda(\theta) + \epsilon([aN] - k)} \sum_{j \in \Gamma(l, k, N)} e^{-Nj \epsilon} \quad (30)$$

Taking the limsup of the log of both sides, we have

$$\lim_{N \to \infty} \frac{1}{\beta N} \log Pr(Q_{[aN] - k} > \beta N) \leq \lim_{N \to \infty} -\theta + \frac{[aN] \Lambda(\theta) + \epsilon}{\beta N} \quad (31)$$

However, from (29) we have

$$\lim_{N \to \infty} \frac{1}{\beta N} \log Pr(Q_{[aN] - k} > \beta N) \leq \left(1 - \frac{aR}{\beta}\right) \sup_{\theta > 0} \{\Lambda(\theta) < \theta R\} + \frac{\epsilon a}{\beta} \quad (32)$$

Next, we find a matching lower bound. Let $\gamma > 0$ be any positive constant. Define

$$\Theta = (b-aR)[aN] + \frac{(b-aR)}{\gamma} N([\gamma] - \gamma) - k(R + \frac{(b-aR)}{\gamma})$$

Let $\Phi = \{N : N > k(R + \frac{aR}{R-aR} + \frac{1}{2})\}$. From (27) and the fact that $\Theta > 0$ for all $N \in \Phi$, we have

$$Pr(Q_{[aN] - k, N} > \beta N) \geq Pr(S_{N[\gamma]} + [aN] - k - RN[\gamma] > \beta N) \geq Pr(S_{N[\gamma]} + [aN] - k - RN[\gamma] > \beta N + \Theta) \geq \frac{S_{N[\gamma]} + [aN] - k}{N[\gamma] + [aN] - k} > R + \frac{\beta - aR}{\gamma}.$$ 

Now, noticing that $\frac{1}{\beta N} \geq \frac{\gamma}{\gamma N[\gamma] + [aN] - k}$ for large enough $N$,

$$\liminf_{N \to \infty} -\frac{1}{\beta N} \log Pr(Q_{[aN] - k} > \beta N) \geq \liminf_{N \to \infty} \frac{\gamma}{\beta N}[\gamma] + [aN] - k - k(N[\gamma] + [aN] - k) \geq \left(1 - \frac{aR}{\beta}\right) \sup_{\theta > 0} \{\Lambda(\theta) < \theta R\} = \left(1 - \frac{aR}{\beta}\right) I(1)$$

Using Cramer’s theorem and optimizing over $\gamma'$, we have

$$\lim_{N \to \infty} \frac{1}{\beta N} \log Pr(Q_{[aN] - k, N} > \beta N) \geq (1 - \frac{aR}{\beta}) \frac{1}{\gamma'} \sup_{\theta > 0} \{\Lambda(\theta) < \theta R\} = \left(1 - \frac{aR}{\beta}\right) I(1)$$

where the first equality follows from Fact 2 (Lemma 3.4 in [2])

**Fact 2:** $\sup_{\theta > 0} \{\Lambda(\theta) < \theta R\} = \inf_{\gamma' \in \mathbb{R}^+} \gamma' \Lambda^*(R + \frac{1}{\gamma'})$. Given the matching lower and upper bounds in (32) and (33) we have the assertion of the lemma. ■

REFERENCES