On the Responsiveness-Diversity-Multiplexing Tradeoff

Petros Elia, Somsak Kittipiyakul, and Tara Javidi

***Abstract*— The work analyzes the error performance of high-SNR, high-rate, point-to-point, outage-limited, wireless communications, with bursty and delay-limited information. In this setting, the bit-arrival process is stochastic and bursty, and the bits are limited by a strict delay condition. In the presence of fixed transmission rate, errors are due to both decoding and delay violations, and are attributed to sequences of atypical bursts of information and atypical fading realizations.**

For the case of fast Rayleigh channel fading, and compound Poisson bit-arrival process, the work presents bounds on a tradeoff between diversity and the ratio of average bit-arrival rate to ergodic-capacity. This tradeoff describes a uniform scalar effect of burstiness on the maximum amount of diversity that can be accumulated given a delay limitation. For large burstiness, the bounds are tight. As a practical consequence, the tradeoff addresses the question of how much of the maximum allowable time should be spent on coding and how much for timely and responsive accommodation of the queue.

I. Introduction

We study the error behavior in point-to-point communications of high-rate, stochastic, bursty and delay-limited traffic, over high-SNR (signal-to-noise ratio), outage-limited, fast-fading channels. In this setting, a bit is in error either when it is decoded incorrectly or when it violates a strict delay limitation, imposed by delay-sensitive applications. Here delay is defined as the time interval between the moment the bit arrives at the source and the moment it is decoded at the receiver. Consequently, in a stable regime of traffic where bits arrive at a long term average rate that is less than the ergodic capacity of the channel, an error occurs either when a sequence of uncommonly ill channels has forced erroneous decoding, or when a sequence of uncommonly large amounts of information has caused large accumulations of bits in the queue. Such error behavior is then a function of the channel statistics, the bit-arrival process statistics, and the delay limitation. Intuitively, more burstiness and unpredictability in the bit arrival process, translates to more delay violations.

This brings to the fore a duality involving the choice of coding duration, and transmission rate. When considering the above outage-limited, bursty, delay-limited system, there are two degrees of freedom that impact error performance. The first degree of freedom is the ratio between rate and ergodic-capacity, and the second is the duration of the coding blocks. In the case of the coding duration, bits are delayed in order to be transmitted over more fading realizations, thus accumulating more temporal diversity, resulting in fewer decoding errors while more delay violations. In the case of the ratio between rate and ergodic-capacity, a larger ratio introduces more decoding errors, but a smaller queueing delay and a decreasing probability of delay violation.

The authors are with the ECE Department, University of California - San Diego, La Jolla, CA, 92093

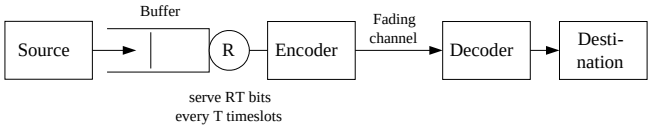

Fig. 1. *Model under consideration: Stochastic-bursty bit-arrival process, outage-limited communications over a fast-fading channel, unlimited storage buffer, first-come first-serve queue service, CSIR, constant transmission/queue-service rate, bit errors due to delay violations and erroneous decoding.*

From the channel point of view, in order to explore the roles of the rate-to-ergodic-capacity ratio and of the coding duration, we adopt the approach of Zheng and Tse [1] who deviate from analyzing capacity regions, and instead adopt a large-deviation approach on analyzing outage events. This approach equates the probability of decoding error to the probability of the rare event that a fading causes an error in the regime of asymptotically high SNR values ρ, i.e., in the limit as $\rho \to \infty$. In this outage limited setting, the rate-to-ergodic-capacity ratio takes the form of the transmission *multiplexing gain* [1]

$$r = \lim_{\rho \to \infty} \big(\text{Rate of Transmission}\big) / \log \rho,$$

with

$$C \approx \log \rho$$

being the high-SNR approximation of the ergodic capacity for the Rayleigh fading single-input, single-output (SISO) channel. Given constant and deterministic bit-arrivals, the asymptotic approximation of the probability of codeword error, is directly implied from [1] to be

$$P_{\text{channel}}(r, T) \approx \rho^{-T(1-r)},$$

revealing the role of the ratio r, and the role of the number of fading realizations T over which coding takes place, in defining the *diversity gain* $T(1-r)$.

Moving onto the point of view of delay violations, we are interested in the probability that, given a stochastic and bursty bit-arrival process, bits violate a *specific* delay limitation, D. For this reason we deviate from the approach of analyzing average delay, and instead employ large-deviations techniques to approximate the probability of the rare event that a sequence of past arrivals with uncommonly large amounts of information, causes a delay violation.

We will show in this paper that, under some traffic models, this probability can be expressed in the form

$$\begin{aligned} P_{\text{delay}}(r,T) &\approx e^{-r\log\rho\cdot F(\text{source statistics},D,T)} \\ &\approx \rho^{-r\cdot F(\text{source statistics},D,T)}, \end{aligned}$$

for some function F. The above two approaches for analyzing errors due to channel and queue dynamics, are based on the principle that rare events occur in the most likely manner. The numerical interpretation of the same principle will then allow that for a choice of r, T, the overall probability of error $P_{\text{total}}(r,T)$, due to both delay and channel, takes the form

$$P_{\text{total}}(r,T) \approx \rho^{-\min\left\{r\cdot F(\text{source statistics},D,T),T(1-r)\right\}}.$$

Optimization over r, T will then provide for the optimal error performance. This will be described in terms of a tradeoff on the maximum amount of diversity gained,

$$\frac{d^*}{D} := -\frac{1}{D}\lim_{\rho\to\infty}\max_{r,T}\frac{\log(P_{\text{total}}(r,T))}{\log\rho}$$

normalized by the amount of diversity that would have been accumulated had there been no burstiness and no delay limitations.

The main result in this work is presented in Theorem 1, in the form of bounds on this optimal tradeoff. As it turns out, in the limit of $\rho\to\infty$, for Rayleigh fast-fading, compound Poisson arrival process governing the random number $\{A_t^\rho\}$ of bits that arrive at discrete time t with average rate

$$\mathbb{E}[A_t^\rho] = \lambda\log\rho,$$

and burstiness measure

$$\frac{1}{\mu} = \frac{1}{2}\frac{\mathbb{E}[(A_t^\rho)^2]}{\mathbb{E}[A_t^\rho]},$$

the tradeoff takes the form

$$\frac{\mu}{(\sqrt{3\mu}+1)^2}\ (1-\lambda) \le \frac{d^*(\lambda)}{D} \le \frac{\mu}{(\sqrt{\mu}+1)^2}\ (1-\lambda).$$

The bounds describe the effect of burstiness and of average throughput on the rate with which diversity is accumulated over time. For large burstiness (small μ), the bounds become tight and the optimal performance takes the form

$$\frac{d^*(\lambda)}{D} \approx \mu\ (1-\lambda).$$

Remark 1: Our aim is to provide simple expressions that give insight on the effect of λ and μ on the maximum amount of time-diversity that can be gained given a specific delay limitation. The corresponding derivations are facilitated in the limit of large D[1]. Despite the fact that D and, as it turns out, the relevant values for T, are all large, our interest remains on the outage-limited regime and on the roles of r and of the ratio D/T. In other words, rather than particular values of D and T, we are interested in the ratio D/T, which is finite and can be small. With this ratio, we address the question of how much of the maximum allowable time should be spent for coding, and how much for accommodating the queue. The same approach also describes another finite and potentially small ratio, between the maximum diversity order that can be accumulated in the presence of burstiness, and the total diversity that would be accumulated in the absence of burstiness and delay limitations.

The remainder of the paper is organized as follows. In Section II we provide a precise characterization of our setting, including the detailed description of the channel model and the bit-arrival process. The statement of the main result is given in Section III, Theorem 1. Section III is then dedicated to the proof of Theorem 1, through analysis of the effects of r, T on the queue dynamics, and description of the optimization over r, T. Other proofs are relegated to the Appendix.

II. System Model

A. Basic setup

We consider a system where bits arrive at the transmitter in a stochastic and bursty manner, independently over time, and where the bit-arrival process can take an unlimited number of states. Upon arrival, the bits are queued in an unlimited storage buffer. The queue size is unknown to all, with the only exception being that the transmitter knows if there exist enough bits to encode or not. Transmission utilizes codes of fixed length T, and average transmission rate R, in the sense that each codeword carries RT bits. The channel experiences fast-fading, which can achieve an unlimited number of states. The statistics of the channel are known to both the transmitter and the receiver, but knowledge of the channel state information, i.e., of the fading realizations, is limited only to the receiver (CSIR).

A bit is considered to be in error when it is incorrectly decoded at the receiver or when the time-duration between the moment the bit arrives at the queue, to the moment the bit is decoded at the receiver, is larger than some delay limitation D. It is assumed that the queue is operating under a steady-state statistical distribution.

B. Model description and assumptions

Discrete time model: We adopt the discrete-time model for both the bit-arrival process and the transmission process. With little loss of generality, we equate the basic units of time for these two processes. This common unit of time will be the *time-slot*.

SISO, i.i.d. Rayleigh, fast-fading channel: Consider the Rayleigh fast-fading channel, with one transmit and one receive antenna, with coding taking place over T time-slots. The $(1\times T)$ received signal vector $\underline{y}$ is given by

$$\underline{y} = [h_1x_1 \quad h_2x_2 \quad ... \quad h_Tx_T] + [w_1 \quad w_2 \quad \cdots \quad w_T]$$

where x_t is the transmitted signal at time t, with $t\in[1,\cdots,T]$. The $(1\times T)$ code vector $\underline{x} := [x_1\ x_2\ ...\ x_T]$ is drawn from a code $\mathcal{X}$. The fading realization h_t and the additive noise w_t at time t, are assumed to be i.i.d., circularly symmetric, complex Gaussian $\mathbb{C}\mathcal{N}(0,1)$ random variables with density function

$$p(u) = \frac{1}{\pi}e^{-|u|^2}.$$

[1]We note that neither D nor T, scale with $\log\rho$.

The value of the fading coefficient will be considered to be entirely known at the receiver, while not known at the transmitter. To ensure the rate requirement, the code has cardinality

$$|\mathcal{X}| = 2^{RT} = \rho^{rT},$$

and to ensure the energy constraint, it is required that

$$|\underline{x}|^2 \ \leq \ \rho T, \quad \text{all } \underline{x} \in \mathcal{X}.$$

We note that the assumption of a common time-scale for bit-arrivals, transmissions and channel coherence, can be readily removed by introducing simple constants which propagate through the analysis. We avoid this for the sake of clarity of exposition. Our results readily accommodate for such constants.

First-come first-serve queue service, with constant rate and fully-diverse coding: We consider the case where service of the queue occurs every T time-slots and where transmission employs a code of rate R and length T, having each bit be transmitted over all fading coefficients. We restrict our attention to the first-come first-serve protocol of queue service. Consequently every T time-slots, the RT oldest bits are instantaneously removed from the queue and are transmitted over the next T time-slots. If not enough data exists in the buffer, null bits are used and the rate is maintained. It is noted that the first-come first-serve queue-service strategy with constant service rate, does not accommodate for any potential knowledge of the queue length at the transmitter. Furthermore, in the absence of channel state information at the transmitter, only a constant service rate provides for optimal, in the high-SNR regime of interest, decoding error performance.

C. Scaling of the transmission rate and of the probability of decoding error:

With respect to the channel model, we adopt the approach taken in [1], where for a choice of T, the error performance, in the presence of constant and deterministic flow, was measured in the form of the *diversity multiplexing gain tradeoff* (DMT), in the high SNR regime, $\rho \to \infty$. In the spirit of DMT, what is of interest here is a rate R that scales as

$$R = r \log \rho$$

with

$$r := \lim_{\rho \to \infty} \frac{R}{\log \rho}$$

being the multiplexing gain. For the specific channel model adopted here, the work in [1] provides for the asymptotic approximation of the optimal probability of codeword error,

$$P_{\text{channel}}(r, T) \doteq \rho^{-T(1-r)}, \quad 0 \leq r \leq 1 \tag{1}$$

or equivalently for the optimal negative SNR exponent

$$d^*_{\text{channel}}(r) = -\lim_{\rho \to \infty} \frac{\log(P_{\text{channel}}(r, T))}{\log \rho} = T(1-r). \tag{2}$$

The notation $\doteq$ is used to describe exponential equality, i.e., $y \doteq \rho^x$ is equivalent to $\lim_{\rho \to \infty} \frac{\log y}{\log \rho} = x$. Similarly for $\dot{\leq}, \dot{\geq}$.

The above conclussion is derived by analyzing the outage sets $\mathcal{O}$, of all fading vectors $\underline{h}$, that cannot sustain enough mutual information. More specifically, decoding error is defined by the outage region of fading vectors

$$\mathcal{O}(\rho, R, T) = \{\underline{h} \ : \ I(\underline{x}; \underline{y} \mid \underline{h}) < RT\} \subset \mathbb{R}^T,$$

that cannot support the transmitted rate R for a given SNR ρ. In the above, $I(\underline{x}; \underline{y} \mid \underline{h})$ is the maximum mutual information accumulated throughout the T channel uses. In the high-SNR regime, outage is a fundamental error-performance limitation, as it defines the probability of error

$$P_{\text{channel}}(r, T) \doteq P(\underline{h} \in \mathcal{O}),$$

and

$$P(\text{channel error} \mid \underline{h} \in \mathcal{O}) \doteq 1.$$

D. Bit-Arrival Model: Scaling of the Average Bit-Arrival Rate and of Burstiness

Let R_{arr} be the *average* rate with which bits arrive at the queue. Given the scale of interest for the transmission rate, it becomes meaningful to scale R_{arr} as

$$R_{\text{arr}} = \lambda \log \rho,$$

where we define

$$\lambda := \lim_{\rho \to \infty} \frac{R_{\text{arr}}}{\log \rho}$$

to be the measure of how close R_{arr} is to the ergodic capacity of the channel. Tuning of the statistics of the bit-arrival process, will adhere to the above scaling, and will define the degree of bit-arrival burstiness.

We specifically choose the compound Poisson bit-arrival process, to be described immediately, which maps well to settings of practical interest, and which admits simple mathematical characterization.

a) Compound Poisson bit-arrival process: We consider the compound Poisson process with exponential distribution of packet sizes. This process governs the statistics of the random integer A^ρ_t,

$$A^\rho_t = \sum_{i=1}^{N^\rho_t} Y^\rho_{i,t},$$

which describes the number of bits that have arrived at integer time t. Parameter ρ is treated as an index. The arrivals are in the form of an accumulation of N^ρ_t packets, with each packet $j \in \{1, \ldots, N^\rho_t\}$, being of size $Y^\rho_{j,t}$.

Each value N^ρ_t is drawn independently from a Poisson distribution having mean that scales as

$$\mathbb{E}[N^\rho_t] = \nu = \nu_0 (\log \rho)^\epsilon,$$

and each value $Y^\rho_{i,t}$ is drawn independently from an exponential distribution with mean and variance that can scale as

$$\mathbb{E}[Y^\rho_{i,t}] = \frac{1}{\mu} = \sqrt{\text{var}(Y^\rho_{i,t})} = (\log \rho)^{(1-\epsilon)},$$

$0 \leq \epsilon \leq 1$. In this work we focus on the case of $\epsilon = 1$.

The above scaling of ν and $\frac{1}{\mu}$ must adopt to the constraint of

$$R_{\text{arr}} = \mathbb{E}\{A_t^\rho\} = \nu\frac{1}{\mu} =: \lambda \log \rho, \tag{3}$$

where then for fixed values of ν_0 and λ, we have that

$$\frac{1}{\mu} = \frac{\lambda}{\nu_0}.$$

Tuning the above variables will define the effect of burstiness. Parameter μ can take any finite value, where a larger value of $1/\mu$, implies larger packets and more burstiness.

We also adopt the condition

$$\lambda < r < 1$$

in consideration of the fact that having $r \geq 1$ would result in $P_{\text{channel}} \doteq P_{\text{total}} \doteq \rho^0$, and having $\lambda \geq r$ would result in $P_{\text{delay}} \doteq P_{\text{total}} \doteq \rho^0$.

A critical function that will play a role in analyzing the probability of delay violation, is the log-moment generating function $\Lambda(\theta)$ of process $\{A_t^\rho\}$. Specifically the function is defined as

$$\Lambda(\theta) := \log \mathbb{E}[e^{\theta A_t}], \quad \theta \in \mathbb{R} \tag{4}$$

where for the compound Poisson bit-arrival process, it takes the form

$$\Lambda(\theta) = \frac{\nu\theta}{\mu - \theta}, \qquad \theta \leq \mu. \tag{5}$$

For a queue with a compound Poisson bit-arrival process, with service rate $R = r \log \rho$, and with both the arrivals and the service occurring every time-slot, then $\Lambda(\theta)$ defines a parameter δ to be the solution to

$$\Lambda(\delta) = R\delta, \tag{6}$$

which in our case, through (3,5,6) takes the form

$$\delta = \mu(r - \lambda)/r > 0. \tag{7}$$

III. Towards the Responsiveness-Diversity-Multiplexing Tradeoff

We now provide simple expressions which, under the constraints of our assumptions, describe bounds on the negative SNR exponent

$$d^*(\lambda) := -\lim_{\rho\to\infty} \max_{r,T} \frac{\log(P_{\text{total}}(\lambda))}{\log \rho},$$

for the optimal total probability of bit error

$$P^*_{\text{total}}(\lambda) \doteq \rho^{-d^*(\lambda)}$$

due to both delay violations and erroneous decoding.

Theorem 1: The optimal tradeoff curve $d^*(\lambda)/D$ is bounded as:

$$\frac{\mu}{(\sqrt{3\mu}+1)^2}\,(1-\lambda) \;\leq\; \frac{d^*(\lambda)}{D} \;\leq\; \frac{\mu}{(\sqrt{\mu}+1)^2}\,(1-\lambda).$$

The above lower and upper bounds are respectively achieved with coding durations T_1, T_2 where $T_1/D = \frac{1}{3}\frac{\sqrt{3\mu}}{\sqrt{3\mu}+1}$ and $T_2/D = \frac{\sqrt{\mu}}{\sqrt{\mu}+1}$.

The derived bounds describe the effect of burstiness and delay-limitation to be a uniform and scalar reduction in diversity, as compared to the DMT expression in (2) that relates to the case of constant and deterministic bit-arrivals, no delay-limitations, and $D = T, r = \lambda$.

The result also indicates that the maximum allowable λ is not reduced, i.e.,

$$\lambda_{\max} := \max_{\lambda\geq 0}\{\lambda \;:\; d^*(\lambda) \geq 0\} = 1 \tag{8}$$

$$= \max_{r\geq 0}\{r \;:\; d^*_{\text{channel}}(r) \geq 0\}. \tag{9}$$

In words, we see that the effects of burstiness and delay limitations are confined to the diversity gained, and do not relate to the maximum value of average bit-arrival rate that can sustain non-catastrophic communication.

The bounds also suggest that in the limit of very low burstiness (μ is large), the effect of delay limitation might persist, while the effect of burstiness vanishes. Finally, in the limit of very high burstiness (μ approaches zero), the bounds are tight and take the form:

$$\frac{d^*(\lambda)}{D} \approx \mu\,(1-\lambda).$$

As a step towards the proof of Theorem 1, we now describe the effect of r, T on the queue dynamics and on the probability of delay violation $P_{\text{delay}}(\lambda, r, T)$.

A. Asymptotic Probability of Delay Violation: Coding Effect

We seek to establish bounds that describe the probability of delay violation $P_{\text{delay}}(\lambda, r, T)$ for a given choice of r, T. We recall that the queue is served every T time-slots with an instantaneous removal of the oldest RT bits. More specifically, after dropping index ρ, we let A_t be the number of bits that arrive in the open interval $(t-1, t)$, and Q_t be the amount of bits in the queue, at the very beginning of time-slot t. This amount of work Q_t also includes the bit of interest. With the queue being served exactly at times tT, the queue size is governed by the following dynamics:

$$Q_t = \begin{cases} [Q_{t-1} + A_t - TR]^+ & \text{if } t = mT,\ m \in \mathbb{Z}, \\ Q_{t-1} + A_t & \text{otherwise,} \end{cases} \tag{10}$$

with $[x]^+ := \max\{x, 0\}$. A standard assumption is that $Q_{mT} = 0$ for integer $m \to -\infty$. Since the arrivals are stationary, we have that for each $i \in \{0, 1, \ldots, T-1\}$, the random sequence $\{Q_{mT+i}\}$ reaches a steady-state random variable Q_i.

The following lemma provides bounds on $P_{\text{delay}}(\lambda, r, T)$, for the arrival-process and queue-service described. Bounds will relate to sets $\mathcal{Q}_{t_i}$ of sample paths that force delay violations. More specifically, $\mathcal{Q}_{t_i}$ will define the set of all error-producing sequences of past arrivals $\omega = \{A_t\}_{-\infty}^{t_i}$, reaching up to a specific time t_i that the bit arrives. The choice of r and T will

define $\mathcal{Q}_{t_i}$ which will in turn define the delay-error behavior through

$$P(\text{delay} \mid \omega \in \mathcal{Q}_{t_i}) = 1$$

and

$$P_{\text{delay}}(\lambda, r, T) = P(\omega \in \mathcal{Q}_{t_i}).$$

Lemma 2: Let $T < D$ and $r \in (\lambda, 1)$. Let $k := D - \lfloor \frac{D}{T} \rfloor T$. Then the probability of delay violation is bounded as

$$\rho^{-\mu(r-\lambda)[D-T-k]^+} \dot{\leq} P_{\text{delay}}(\lambda, r, T) \dot{\leq} \rho^{-\mu(r-\lambda)[D-2T-k]^+} \tag{11}$$

and more loosely bounded as

$$\rho^{-\mu(r-\lambda)(D-T)} \dot{\leq} P_{\text{delay}}(\lambda, r, T) \dot{\leq} \rho^{-\mu(r-\lambda)[D-3T]^+}. \tag{12}$$

Proof of Lemma 2:

We focus on bits that arrive at some arbitrary time $t = mT+i$, for some finite negative integer m, where $i \in [0, T-1]$. Without loss of generality, and for notational simplicity, we set $m = 0$ and thus consider the arrival of the above bits to take place at time i. Furthermore it is easy to show that in our setting and our scale of interest, the average probability of delay for bits that arrive at time i, is equal to the probability of delay for the last bit that arrives at the same time i. Consequently we focus on the last bit that arrives at this time i.

We will explore the different cases that arise from having different values of $D + i$ and T, and will explain how a combination of i, D, T define the sample paths ω that force error. Towards this, we have

$$\begin{aligned}
\mathcal{Q}_i &= \left\{\omega \ : \ i + (T-i) + \lceil \frac{Q_i(\omega)}{RT} \rceil T > \lfloor \frac{D+i}{T} \rfloor T \right\} \\
&= \left\{\omega \ : \ \lceil \frac{Q_i(\omega)}{RT} \rceil T > \lfloor \frac{D+i}{T} \rfloor T - T \right\} \\
&= \left\{\omega \ : \ \lceil \frac{Q_i(\omega)}{RT} \rceil T > \lfloor \frac{D+i-T}{T} \rfloor T \right\} \\
&= \left\{\omega \ : \ \lceil \frac{Q_i(\omega)}{RT} \rceil > \lfloor \frac{D+i-T}{T} \rfloor \right\} \\
&= \left\{\omega \ : \ \frac{Q_i(\omega)}{RT} > \lfloor \frac{D+i-T}{T} \rfloor \right\},
\end{aligned}$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are respectively the floor and ceiling functions.

To clarify the above, we describe the meaning of each term.

$$\mathcal{O}_i = \left\{\omega \ : \ \underbrace{i}_{E_1} + \underbrace{(T-i)}_{E_2} + \overbrace{\underbrace{\lceil \frac{Q_i(\omega)}{RT} \rceil}_{E_3} T}^{E_4} > \overbrace{\lfloor \frac{D+i}{T} \rfloor T}^{E_5} \right\}.$$

E_1: bit arrives at normalized time index $E_1 = i$.

E_2: number of time-slots that the bit waits for the first block of TR bits to leave the queue.

E_3: the index of the block that delivers the bit.

E_4: total time spent for servicing the above E_3 blocks. This is the duration between the beginning of the first service, to the completion of decoding for the bit. Note that E_4 accounts for the coding duration.

E_5: The last decoding opportunity for the bit.

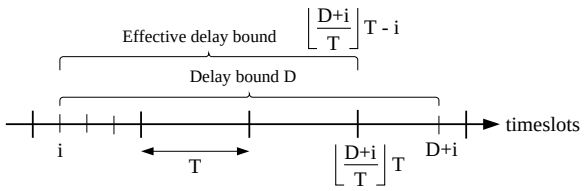

Fig. 2. Timing diagram indicating the different intervals in the lifespan of a bit that has arrived last, at time i.

We continue with the proof and let $n := \lfloor \frac{D}{T} \rfloor$ and recall that $k = D - nT = D - \lfloor \frac{D}{T} \rfloor T$, $k \in \{0, 1, \ldots, T-1\}$. Then

$$\left\lfloor \frac{D-T+i}{T} \right\rfloor = \begin{cases} n-1, & \text{if } 0 \leq i \leq T-k-1, \\ n, & \text{if } T-k \leq i \leq T-1. \end{cases}$$

Consequently,

$$\mathcal{Q}_i = \begin{cases} \{\omega \ : \ Q_i(\omega) > (n-1)(RT), & i \in [0, T-k-1]\} \\ \{\omega \ : \ Q_i(\omega) > n(RT), & i \in [T-k, T-1]\}. \end{cases}$$

With $P(\omega \in \mathcal{Q}_i)$ being the probability of delay violation for bits that arrive at time i, and for i uniformly distributed in $[0, T-1]$, we have that

$$\begin{aligned}
&P_{\text{delay}}(\lambda, r, T) \\
&= \frac{1}{T} \sum_{i=0}^{T-1} P(\omega \in \mathcal{Q}_i) \\
&\doteq \sum_{i=0}^{T-1} P(\omega \in \mathcal{Q}_i), \qquad (\text{since } T \doteq \rho^0) \\
&= \sum_{i=0}^{T-k-1} P(\omega \in \mathcal{Q}_i) + \sum_{i=T-k}^{T-1} P(\omega \in \mathcal{Q}_i) \\
&\doteq \max\{P(\omega \in \mathcal{Q}_{T-k-1}), P(\omega \in \mathcal{Q}_{T-1})\} \\
&\doteq \max\{P(Q_{T-k-1} > (n-1)TR), P(Q_{T-1} > nTR)\} \\
&\doteq \max\{P(Q_{T-k-1} > (D-T-k)R), \\
&\qquad\qquad P(Q_{T-1} > (D-k)R)\}
\end{aligned} \tag{13}$$

where to get the fourth equation, we used that

$$\begin{aligned}
\mathcal{Q}_{T-k-1} \supseteq \mathcal{Q}_{T-k-2} \supseteq \cdots \supseteq \mathcal{Q}_0 \\
\mathcal{Q}_{T-1} \supseteq \mathcal{Q}_{T-2} \supseteq \cdots \supseteq \mathcal{Q}_{T-k}
\end{aligned}$$

which holds because for any j, i such that $0 \leq j < i \leq T-1$, it is the case that

$$Q_i(\omega) = Q_j(\omega) + \underbrace{A_{j+1}(\omega) + \ldots + A_i(\omega)}_{\geq 0} \geq Q_j(\omega), \quad \forall \omega.$$

To find the dominant term in (13), we look for the probability for the general form

$$P(Q_i > D'R),$$

by relating the steady-state distributions of Q_0 and Q_i, for any $i \in \{0, 1, \ldots, T-1\}$ and any $D' \geq 2T$. First we easily establish a lower bound:

$$\begin{aligned} & P(Q_i > D'R) \\ & = P(Q_0 + \underbrace{A_1 + \ldots + A_i}_{\geq 0} > D'R) \\ & \geq P(Q_0 > D'R). \end{aligned} \tag{14}$$

Moving onto the upper bound for $P(Q_i > D'R)$, we first recall from (10) that

$$Q_T(\omega) = [Q_i(\omega) + A_{i+1}(\omega) + \cdots + A_T(\omega) - TR]^+, \quad \forall \omega.$$

We are only interested in sample paths for which $Q_i(\omega) > DR'$. For such sample paths, the following sequence holds:

$$\begin{aligned} & Q_i(\omega) > D'R \\ & \Rightarrow Q_i(\omega) - TR > 0, \quad \text{since} \quad D' \geq 2T \\ & \Rightarrow Q_i(\omega) - TR + \underbrace{A_{i+1}(\omega) + \cdots + A_T(\omega)}_{\geq 0} > 0 \end{aligned}$$

$$\Rightarrow Q_T(\omega) = Q_i(\omega) + A_{i+1}(\omega) + \cdots + A_T(\omega) - TR$$

$$\begin{aligned} & \Rightarrow Q_i(\omega) = Q_T(\omega) - (A_{i+1}(\omega) + \cdots + A_T(\omega)) + TR \\ & \Rightarrow P(Q_i > D'R) \\ & = P(Q_T - \{A_{i+1} + \ldots + A_T\} + TR > D'R) \\ & \leq P(Q_T > (D' - T)R). \end{aligned}$$

Due to stationarity of both the arrivals and service at times that are finite multiples of T, we have:

$$P(Q_i > D'R) \leq P(Q_0 > (D' - T)R). \tag{15}$$

In conjunction with the lower bound in (14), we have

$$P(Q_0 > D'R) \leq P(Q_i > D'R) \leq P(Q_0 > (D' - T)R). \tag{16}$$

We can now establish the dominant term in (13) by substituting for i and for the two cases of D', to get two sets of bounds:

$$\begin{aligned} & P(Q_0 < \overbrace{(D - T - k)}^{D' \text{: case 1}} R) \leq \\ & \qquad P(Q_{T-k-1} > (D - T - k)R) \leq \\ & \qquad\qquad P(Q_0 < (D - 2T - k)R), \end{aligned} \tag{17}$$

$$\begin{aligned} & P(Q_0 < \overbrace{(D - k)}^{D' \text{: case 2}} R) \leq \\ & \qquad P(Q_{T-1} > (D - k)R) \leq \\ & \qquad\qquad P(Q_0 < (D - T - k)R) \end{aligned}$$

which jointly imply that

$$P(Q_{T-k-1} > (D - T - k)R) \geq P(Q_{T-1} > (D - k)R).$$

This identifies the dominant term in (13) and the value for

$$P_{\text{delay}}(\lambda, r, T) \doteq P(Q_{T-k-1} > (D - T - k)R). \tag{18}$$

Then the bound from (17) applies to give the following bound on $P_{\text{delay}}(\lambda, r, T)$, in terms of the steady-state Q_0

$$\begin{aligned} & P(Q_0 > (D - T - k)R) \dot{\leq} \\ & \qquad P_{\text{delay}}(\lambda, r, T) \dot{\leq} \\ & \qquad\qquad P(Q_0 > (D - 2T - k)R). \end{aligned} \tag{19}$$

Directly from Claim 1 in the Appendix, we know that

$$P(Q_0 > D'R) \doteq e^{-\delta D' R} = \rho^{-\delta r D'}, \quad \text{for} \quad D' \to \infty.$$

Given that

$$P_{\text{delay}}(Q_0 > D'R \mid D' \leq 0) = 1,$$

we have the above bounds in (19) taking the form

$$\rho^{-\delta r [D-T-k]^+} \dot{\leq} P_{\text{delay}}(\lambda, r, T) \dot{\leq} \rho^{-\delta r [D-2T-k]^+}, \tag{20}$$

where we recall that $\delta = \mu \frac{r-\lambda}{r} > 0$, and satisfies

$$\log \mathbb{E}[e^{\delta A_t}] = R\delta. \tag{21}$$

Consequently we get,

$$\rho^{-\mu(r-\lambda)[D-T-k]^+} \dot{\leq} P_{\text{delay}}(\lambda, r, T) \dot{\leq} \rho^{-\mu(r-\lambda)[D-2T-k]^+}, \tag{22}$$

and the more loose and simpler bound

$$\rho^{-\mu(r-\lambda)(D-T)} \dot{\leq} P_{\text{delay}}(\lambda, r, T) \dot{\leq} \rho^{-\mu(r-\lambda)[D-3T]^+}. \tag{23}$$

□

We now proceed with the conclusion of the proof for Theorem 1.

B. Optimization of the Total Error Probability - Conclusion of Proof for Theorem 1

Having established $P_{\text{delay}}(\lambda, r, T)$, we look to first establish the total error probability $P_{\text{total}}(\lambda, r, T)$, for a given choice of r, T. We will then optimize over all choices of r, T and obtain the bounds on the optimal tradeoff.

Proof of Theorem 1:

The total probability of error $P_{\text{total}}(\lambda, r, T)$, due to both delay violation and channel decoding errors, can be expressed as

$$\begin{aligned} & P_{\text{total}}(\lambda, r, T) = \\ & \quad P_{\text{channel}}(\lambda, r, T) + (1 - P_{\text{channel}}(\lambda, r, T)) P_{\text{delay}}(\lambda, r, T). \end{aligned}$$

We rewrite the bound of Lemma 2, as

$$P_{\text{delay}}(\lambda, r, T) \doteq \rho^{-\mu(r-\lambda)(D - a_T T)}, \tag{24}$$

for some unknown value a_T such that $a_T \in [a_{\min}, a_{\max}]$, where $a_{\min} = 1, a_{\max} = \min\{3, D/T\}$ and where $\lambda < r < 1$ and $0 < T < D$. Using the value of $P_{\text{channel}}(\lambda, r, T)$ from (1)

and the new form for $P_{\text{delay}}(\lambda, r, T)$ from (24), we get that in the scale of interest[2]:

$$\begin{aligned} P_{\text{total}}(\lambda, r, T) &\doteq P_{\text{channel}}(\lambda, r, T) + P_{\text{delay}}(\lambda, r, T) \\ &\doteq \rho^{-\mu(r-\lambda)(D-a_T T)} + \rho^{-T(1-r)}. \\ &\doteq \rho^{-\min\{\mu(r-\lambda)(D-a_T T), T(1-r)\}}. \end{aligned}$$

Let

$$\begin{aligned} \xi(r, T) &:= -\lim_{\rho\to\infty} \frac{\log P_{\text{total}}(r, T)}{\log \rho} \\ &= \min\{\mu(r-\lambda)(D-a_T T), T(1-r)\} \end{aligned} \quad (25)$$

and let

$$\begin{aligned} d^*(\lambda) &:= d^*(\lambda, a_T) \\ &= \max_{r,T} \min\{\mu(r-\lambda)(D-a_T T), T(1-r)\} \end{aligned} \quad (26)$$

such that $\lambda < r < 1$ and $0 < T < D$. Since this is a double maximization problem, we have

$$\begin{aligned} &d^*(\lambda, a_T) \\ &= \max_{0<T<D} \left\{ \max_{\lambda<r<1} \min\{\mu(r-\lambda)(D-a_T T), T(1-r)\} \right\}. \end{aligned} \quad (27)$$

Fix T and consider the maximization over r, to find

$$\xi(r^*(T), T) = \max_{\lambda<r<1} \min\{\mu(r-\lambda)(D-a_T T), T(1-r)\} \quad (28)$$

for some optimizing value $r^*(T)$. The fact that $\mu(r-\lambda)(D-a_T T)$ is an increasing function on r, and $T(1-r)$ a decreasing function on r, makes it so that $\min\{\mu(r-\lambda)(D-a_T T), T(1-r)\}$ is maximized[3] when

$$\delta r^*(T)(D-a_T T) = T(1-r^*(T)). \quad (29)$$

Solving for $r^*(T)$ we get the optimal multiplexing gain

$$r^*(T) = 1 - \frac{1-\lambda}{1+\frac{T}{\mu(D-a_T T)}} \quad (30)$$

for a given T. Since $0 < \lambda < 1$ and $0 < T < D$, the condition $\lambda < r^*(T) < 1$ is satisfied[4].

Combining (28) and (30) gives the optimal exponent, for a specific choice of T, to be

$$\xi(r^*(T), T) = T(1-r^*(T)) = \frac{T(1-\lambda)}{1+\frac{T}{\mu(D-a_T T)}}. \quad (31)$$

The next step is to optimize over T and establish

$$d^*(\lambda) = d^*(\lambda, a_T) = \max_{0<T<D} \xi(r^*(T), T). \quad (32)$$

[2]It is easy to see that in the scale of interest, the probability of bit error is of the same exponential order as the probability of codeword error. Furthermore, the fact that there exist codes that can accumulate the described diversity, has been established in [1], and [12]- [14].

[3]The maximization is directly valid since the two terms are independent of ρ.

[4]The condition $\lambda < r^*(T) < 1$ is also satisfied because if this were not true, which would happen either when $r^*(T) < \lambda$ or when $r^*(T) > 1$, then one of the terms in (28) would be zero while the other term would be strictly positive. This contradicts the fact that $r^*(T)$ equates the two terms in (29).

Since we do not know the exact value of a_T but only know that $a_T \in [a_{\min}, a_{\max}]$, we are limited to providing bounds on $d^*(\lambda)$.

For this, set $a \in [a_{\min}, a_{\max}]$, let

$$f(a, T) := \frac{\xi(r^*(T), T)}{1-\lambda} = \frac{T}{1+\frac{T}{\mu(D-aT)}}, \quad (33)$$

and note that since $f(a, T)$ is monotonically decreasing on a, it is the case that

$$f(a_{\max}, T) \le f(a_T, T) \le f(a_{\min}, T), \quad (34)$$

for any $T \in (0, D)$. Consequently

$$\max_{0<T<D} f(a_{\max}, T) \le \underbrace{\max_{0<T<D} f(a_T, T)}_{d^*(\lambda)/(1-\lambda)} \le \max_{0<T<D} f(a_{\min}, T). \quad (35)$$

Our goal now is to evaluate

$$\max_{0<T<D} f(a_{\max}, T) \quad \text{and} \quad \max_{0<T<D} f(a_{\min}, T).$$

For this we note that

$$\max_{0<T<D} f(a_{\max}, T) = \max_{0<T\le D/3} f(3, T) \quad (36)$$

because

$$f(a_{\max}, T) = 0, \quad D/3 \le T < D.$$

As a result, the bounds take the form

$$\max_{0<T<D/3} f(3, T) \le \underbrace{\max_{0<T<D} f(a_T, T)}_{d^*(\lambda)/(1-\lambda)} \le \max_{0<T<D} f(1, T). \quad (37)$$

To evaluate the above, we look to find the optimizer

$$T_a^* = \arg \max_{0<T<D/a} f(a, T)$$

for the two distinct cases of $a = 1$ and $a = 3$. For this we differentiate $f(a, T)$ with respect to T, and solve for $f'(a, T_a^*) = 0$. In our case, this gives

$$T_a^* = \frac{D}{a}\left(\frac{\sqrt{\mu a}}{\sqrt{\mu a}+1}\right). \quad (38)$$

Using this value of T_a^*, results in

$$\max_{0<T<D/a} f(a, T) = f(a, T_a^*) = \frac{\mu}{(\sqrt{\mu a}+1)^2} D. \quad (39)$$

Using the bounds from (37) results in

$$\frac{\mu}{(\sqrt{3\mu}+1)^2}(1-\lambda) \le \frac{d^*(\lambda)}{D} \le \frac{\mu}{(\sqrt{\mu}+1)^2}(1-\lambda). \quad (40)$$

This concludes the proof of Theorem 1. □

IV. Conclusion

In this work we derived simple expressions that give insight on the effect that bit-arrival burstiness has on the maximum amount of time-diversity that can be gained, under a given delay limitation.

The expressions, in the form of bounds, describe a tradeoff between diversity and the ratio of average bit-arrival rate to ergodic-capacity. As a practical consequence, the tradeoff tells us how to better balance the effects of channel-atypicality (outage) and burstiness-atypicality, by proper choice of information rate and encoding duration. The tradeoff addresses the question of how much of the maximum allowable time should be spent on coding and how much for timely accommodation of the queue.

V. Appendix

Claim 1: Consider a batch queueing system where arrivals happen every time-slot t, t an integer, and where the number of bits A_t that arrive at time t, follow the i.i.d. compound Poisson process. Let the queue be served on a first-come first-serve basis by instantaneously removing RT bits, at times mT, m an integer. Let the queue be stable, i.e., let $\mathbb{E}[A_0] < R$. Let Q_t be the queue length (in bits) at time t, and let $Q_{-t} \equiv 0$ for $t \to \infty$, allowing Q_0 to be governed by the steady-state distribution. Then,

$$\lim_{D' \to \infty} \frac{1}{D'} \log P(Q_0 > D'R) = -R\delta \tag{41}$$

where $\delta > 0$ is defined as in (6) to be such that $\Lambda(\delta) = R\delta$.

Proof of Claim 1

We consider a new system with a normalized time-scale, having a time-slot being equal to T time-slots of the original system described in the claim. Consequently, this new system is defined by a new i.i.d. arrival process $\{\hat{A}_t\}$, where

$$\hat{A}_t = A_{tT+1} + \cdots + A_{(t+1)T}$$

defines the number of bits that arrive during (new) time-slot t. The new system is also defined by the service that happens every (new) time-slot with fixed rate

$$\hat{R} = RT, \quad \text{bits per (new) time-slot.}$$

The new queue length $\hat{Q}_t(\omega)$ satisfies

$$\hat{Q}_t(\omega) = Q_{tT}(\omega), \quad \forall \omega$$

and follows a new set of queue dynamics

$$\hat{Q}_{t+1} = (\hat{Q}_t + \hat{A}_t - \hat{R})^+. \tag{42}$$

Given the standard assumption that for every sample path ω, $Q_{-tT}(\omega) = \hat{Q}_{-t}(\omega) = 0$, for $t \to \infty$, we have a common steady-state distribution Q_0 for $t = 0$. Consequently, Theorem 1.4 from [19] applies, to give

$$\lim_{D' \to \infty} \frac{1}{D'} \log P(Q_0 > D'R) = -\delta_{RT} R \tag{43}$$

where δ_{RT} is the solution to

$$\hat{\Lambda}(\delta_{RT}) = \delta_{RT} RT$$

and where $\hat{\Lambda}(\delta_{RT})$ is the log-moment generating function for random variable $\hat{A}_0$. To conclude the proof, we simply have to show that $\delta_{RT} = \delta$. To do so, we notice that for any $\theta \in \mathbb{R}$, then

$$\begin{aligned} \hat{\Lambda}(\theta) = \log E e^{\theta(\hat{A}_0)} &= \log E e^{\theta(A_1 + \cdots + A_T)} \\ &= T \log E e^{\theta A_1} = T\Lambda(\theta) \end{aligned}$$

since $A_1, \ldots, A_T$ are i.i.d. By definition of δ_{RT}, we have that

$$\delta_{RT} R = \hat{\Lambda}(\delta_{RT})/T = \Lambda(\delta_{RT})$$

which agrees with the definition of δ. This means that δ and δ_{RT} are the solutions to the same equation. The proof is concluded by noticing that $RT\delta$ is linear in δ, and that $\Lambda(\theta)$ is convex and monotonically increasing. □

References

[1] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Info. Theory*, vol. 49, no. 5, pp. 1073-1096, May 2003.

[2] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424-428, Aug. 1993.

[3] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential approximations for tail probabilities in queues I: waiting times," *Operations Research*, v. 43, no. 5, 1995.

[4] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential approximations for tail probabilities in queues II: sojourn time and workload," *Operations Research*, v. 44, no. 5, 1996.

[5] G. L. Choudhury, D. M. Lucantoni and W. Whitt, "Squeezing the most out of ATM," IEEE Trans. on Communications, v. 44, no.2, Feb 1996.

[6] J. Razavilar, K. J. R. Liu, and S. I. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *IEEE Trans. on Commun.*, vol. 50, no. 3, pp. 484–494, March 2002.

[7] T. Holliday and A. Goldsmith, "Joint source and channel coding for MIMO systems," in *Proc. Allerton Conf. Comm., Control and Computing*, Oct. 2004.

[8] R. Negi and S. Goel, "An information-theoretic approach to queuing in wireless channels with large delay bounds," *Globecomm'04*, 2004.

[9] 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 Comm.*, v. 4, no. 3, May 2005.

[10] T. Holliday and A. Goldsmith, "Optimizing end-to-end distortion in MIMO systems", " *Proc. IEEE Int. Symp. Inform. Th (ISIT 2005).*

[11] S. Kittipiyakul and T. Javidi, "Optimal operating point in MIMO channel for delay-sensitive and bursty traffic," *Proc. IEEE Int. Symp. Inform. Th (ISIT 2006).*

[12] P. Elia, K. Raj Kumar, Sameer A. Pawar, P. Vijay Kumar and Hsiao-feng Lu, "Explicit, minimum-delay space-time codes achieving the diversity-multiplexing gain tradeoff," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, September 2006.

[13] P. Elia *Asymptotic Universal Optimality in Wireless Multi-Antenna Communications and Wireless Networks*, Ph.D. thesis, USC, 2006.

[14] S. Yang and J.-C. Belfiore, "Optimal space-time codes for the amplify-and-forward cooperative channel, " *IEEE Trans. Inform. Theory*, vol. 53, no. 2, February 2007.

[15] F. Kelly, "Notes on effective bandwidths," *Stochastic Networks: Theory and Applications*, vol. 4, 1996.

[16] A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd edition, Springer-Verlag, New York, 1998.

[17] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*, Chapman and Hall, 1995.

[18] F. den Hollander, *Large Deviations*, American Mathematical Society, Providence, RI, 2000.

[19] A. Ganesh, N. O'Connel, and D. Wischik, *Big Queues*, Springer-Verlag, Berlin, 2004.