CHALMERS
UNIVERSITY OF TECHNOLOGY

Delay and Peak-Age Violation Probability in Short-Packet Transmissions

Citation for the original published paper (version of record):
Devassy, R., Durisi, G., Ferrante, G. et al (2018). Delay and Peak-Age Violation Probability in Short-Packet Transmissions. IEEE International Symposium on Information Theory - Proceedings, 2018-June: 2471-2475. http://dx.doi.org/10.1109/ISIT.2018.8437671

Delay and Peak-Age Violation Probability in Short-Packet Transmissions

Rahul Devassy[1], Giuseppe Durisi[1], Guido Carlo Ferrante[1], Osvaldo Simeone[2], and Elif Uysal-Biyikoglu[3]

[1]Chalmers University of Technology, 41296 Gothenburg, Sweden
[2]King's College London, London, WC2R 2LS, United Kingdom
[3]Middle East Technical University, Ankara, Turkey

***Abstract*—This paper investigates the distribution of delay and peak age of information in a communication system where packets, generated according to an independent and identically distributed Bernoulli process, are placed in a single-server queue with first-come first-served discipline and transmitted over an additive white Gaussian noise (AWGN) channel. When a packet is correctly decoded, the sender receives an instantaneous error-free positive acknowledgment, upon which it removes the packet from the buffer. In the case of negative acknowledgment, the packet is retransmitted. By leveraging finite-blocklength results for the AWGN channel, we characterize the delay violation and the peak-age violation probability without resorting to approximations based on large deviation theory as in previous literature. Our analysis reveals that there exists an optimum blocklength that minimizes the delay violation and the peak-age violation probabilities. We also show that one can find two blocklength values that result in very similar average delay but significantly different delay violation probabilities. This highlights the importance of focusing on violation probabilities rather than on averages.**

***Index Terms*—Delay, finite blocklength, age of information, queuing.**

I. Introduction

Emerging wireless applications such as factory automation and vehicular communication require the availability of mission-critical links that are able to deliver short information packets within stringent latency and reliability constraints. As shown in [1], finite-blocklength information theory provides accurate tools to describe the tradeoff between latency, reliability, and rate when transmitting short packets. Leveraging tools from non-asymptotic information theory, the purpose of this paper is to analyze the probability that the delay or the peak age [2, Def. 3] exceeds a predetermined threshold in a point-to-point communication system with random information-packet arrivals per channel use. The analysis assumes a single-server queue operating according to a first-come-first-served (FCFS) policy.

Related Work: Aside from the work by Telatar and Gallager [3], who employed an error-exponent approach, most works in queuing analysis of communication links rely on a bit-pipe abstraction of the physical layer. Accordingly, bits are delivered reliably at a rate equal to the channel capacity, or in the case of fading channels, at a rate equal to the outage capacity for a given outage probability. These works may be classified into three broad categories: (i) analyses of the steady-state average delay; (ii) analyses of the delay violation probability using large deviation theory (see [4], [5] and references therein); and (iii) analyses of throughput-delay tradeoff under deadline constraints [6], [7]. However, the bit-pipe abstraction is not suitable when the latency constraints prevent the use of channel codes with long blocklength. Indeed, outage and ergodic capacity are poor performance benchmarks when packets are short [8], and using them may result in inaccurate delay estimates.

Recognizing these limitations, Hamidi-Sepehr *et al.* [9] analyzed the queuing behavior when BCH codes are used. Specifically, they evaluated both the probability distribution of the steady-state queue size and the average delay. A different approach, which relies on random coding, is to replace capacity or outage capacity with the more accurate second-order asymptotic approximations obtained in [1], [10]. This approach has been used in [11] to study the throughput achievable over a fading channel under a constraint on the probability of buffer overflow; in [12] to analyze the packet delay violation probability in the presence of perfect channel-state information at the transmitter, which allows for rate adaptation; and in [13] to design the downlink of an ultra-reliable transmission system under a constraint on the end-to-end delay. All these works rely on large-deviation theory through effective capacity [14], stochastic network calculus [15], and effective bandwidth [16], hence providing tight delay estimates only in the asymptotic limit of large delay.

For applications in which packets carry status updates, the time elapsed since the newest update available at the destination was generated at the source, commonly referred to as age of information, is more relevant than delay. Most previous analyses of the age of information focus on its average or peak value (see [2] and references therein), and rely on simple physical-layer models. A recent exception is [17], where the stationary distribution of the peak age is characterized, and [18], where generalized age penalty functions are analyzed.

This work was partly funded by the Swedish Research Council under grant 2012-4571. The simulations were performed in part on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE.

Osvaldo Simeone has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

Elif Uysal-Biyikoglu has received funding from TÜBİTAK (Grant No. 117E215).

Contributions: We analyze the delay and peak-age violation probabilities achievable over an additive white Gaussian noise (AWGN) channel where information packets arrive in each channel use (CU) according to an independent and identically distributed (i.i.d.) Bernoulli process and are transmitted using an FCFS policy with automatic repeat request. Our specific contributions are as follows:

- We determine in closed form the probability-generating functions (PGFs) of delay and peak age at steady state. Delay and peak-age violation probabilities can be efficiently obtained from the derived PGFs through an inverse transform. We also present an accurate approximation of this inverse transform based on saddlepoint methods.
- We numerically illustrate the dependence of delay on the blocklength. Specifically, we show that there exist two blocklength values resulting in the same average delay, but yielding delay violation probabilities that differ by two orders of magnitude. This shows that average delay is insufficient in capturing performance.
- Finally, we discuss the accuracy of delay violation estimates based on the large-deviation tools used in [12].

Notation: Uppercase boldface letters denote random quantities and lightface letters denote deterministic quantities. The distribution of a random variable X is denoted by P_{X}. With $\mathbb{E}[\cdot]$ we denote the expectation operator. The indicator function and the ceil function are denoted by $\mathbf{1}\{\cdot\}$ and $\lceil\cdot\rceil$ respectively. We let $\mathrm{Bern}(p)$ denote a Bernoulli-distributed random variable with parameter p, $\mathrm{Binom}(n,p)$ denote a Binomial-distributed random variable with parameters n and p, and $\mathrm{Geom}(p)$ a geometrically distributed random variable with parameter p. The PGF of a non-negative integer-valued random variable X is $G_{\mathsf{X}}(s) = \mathbb{E}\big[s^{\mathsf{X}}\big]$.

II. Frame-Synchronous Model

We consider a point-to-point discrete-time AWGN channel. The information-packet arrival process is i.i.d. Bernoulli over the CUs. Specifically, the probability of a new packet arrival in each CU is λ. The information packets are stored at the transmitter in a single-server queue operating according to an FCFS policy. Each information packet consists of k bits, which are mapped into a codeword of blocklength n CUs and power P.[1] The packet error probability is denoted by $\epsilon > 0$. A packet is removed from the buffer when its reception is acknowledged by the receiver through an ACK feedback. If the codeword is not correctly decoded, the receiver sends a NACK message and the codeword is retransmitted. We assume perfect error detection at the receiver and instantaneous error-free ACK/NACK transmission, as commonly done in the literature.

We will first assume that time is organized into time frames of duration n CUs so that the transmission of a codeword can only start at the beginning of a time frame. Under this assumption, if an information packet arrives when the buffer is empty, its transmission is scheduled for the next available frame. We refer to this setup as being *frame synchronous*. In Section IV, we shall relax this assumption and allow transmission to start in the next available CU when the buffer is empty. We refer to this setup as being *frame asynchronous*. This model yields a reduction in latency at the cost of a more involved frame-synchronization procedure. Under the frame-synchronous assumption, the system can be modeled as a $Geo/G/1$ queue with bulk arrivals, sometimes denoted $Geo^{[X]}/G/1$ (see [19, Sec. 4.6.2]).

We group together all packets arriving within a time frame as a *bulk*, and study the evolution of the transmitter's buffer along the time index t running over the time frames. Let B_t be the number of packets received in the t-th time frame. It follows that the process $\{\mathsf{B}_t\}_{t=1}^{\infty}$ is i.i.d. with $\mathrm{Binom}(n,\lambda)$ marginal distribution. When $\mathsf{B}_t > 0$, we say that a bulk has been received at time frame t. Furthermore, we denote by Q_t the number of bulks remaining in queue at the start of the $(t+1)$th time frame.

Let T_m be the frame index corresponding to the arrival time of the mth bulk, and N_m be the number of packets in the mth bulk. Note that $\{\mathsf{N}_m\}_{m=1}^{\infty}$ is an i.i.d. process with marginal distribution equal to the conditional distribution of B_t given the event $\{\mathsf{B}_t > 0\}$. We denote by W_m the waiting time of the mth bulk, i.e., the number of frames the first packet in the bulk remains in the queue before being served. Moreover, S_m is the service time of the mth bulk, i.e., the total number of frames needed to successfully transmit all packets in the mth bulk. The service process $\{\mathsf{S}_m\}_{m=1}^{\infty}$ is i.i.d. with

$$\mathsf{S}_m \sim \sum_{k=1}^{\mathsf{N}_m} \mathsf{H}_k \tag{1}$$

where the variables $\{\mathsf{H}_k\}_{k=1}^{\mathsf{N}_m}$ are i.i.d. $\mathrm{Geom}(1-\epsilon)$-distributed and independent of N_m. Each variable H_k represents the number of time frames needed to reliably deliver one packet. Finally, the delay $\mathsf{D}_m = \mathsf{W}_m + \mathsf{S}_m$ of the mth bulk (measured in frames) is the sum of waiting time W_m and service time S_m. For this queuing system, the process $\{\mathsf{D}_m\}_{m=1}^{\infty}$ has a steady-state distribution as long as $\lambda n < 1-\epsilon$. This distribution is studied in the next section. We will discuss the peak-age metric in Section V.

III. Steady-State Delay Violation Probability

In this section, we focus on the analysis of the steady-state delay violation probability. This is defined as

$$P_{\mathrm{dv}}(d_0) = \lim_{m\to\infty} \Pr\{\mathsf{D}_m \geq d_0/n\} = \Pr\{\mathsf{D} \geq d_0/n\}, \tag{2}$$

where D is the steady-state delay and d_0 is the desired latency constraint (measured in CUs). To characterize $P_{\mathrm{dv}}(d_0)$, we will first derive the PGF of D, and then obtain $P_{\mathrm{dv}}(d_0)$ implicitly through an inversion formula. As the PGF of the delay D for our setup is not directly available in the literature, although its derivation follows along the steps described in [19, Sec. 4.6.2], we provide it in the following theorem.

Theorem 1: For every $\epsilon > 0$ such that $\lambda n < 1-\epsilon$, the PGF of the steady-state delay D for the frame-synchronous model is

[1] We assume that the variance of the Gaussian additive noise is one. So P is also the signal-to-noise ratio.

$$G_{\mathsf{D}}(s) = (1-\lambda n/(1-\epsilon)) \cdot \frac{(1-s)((1-\lambda)^n(1-\epsilon s)^n - (1-\lambda+(\lambda-\epsilon)s)^n)}{(1-(1-\lambda)^n)(s(1-\epsilon s)^n - (1-\lambda+(\lambda-\epsilon)s)^n)}. \quad (3)$$

Proof: See Appendix A. ∎

The delay violation probability (2) can be obtained from (3) through the following inversion formula

$$P_{\mathrm{dv}}(d_0) = 1 - \left(\frac{1}{2\pi i}\oint_C \frac{G_{\mathsf{D}}(s)}{(1-s)s^{d-1}}\mathrm{d}s\right)\mathbf{1}\{d \geq 2\} \quad (4)$$

where $d = \lceil d_0/n \rceil$ and C is a circle centered at the origin of the complex plane enclosing all poles of $G_{\mathsf{D}}(s)/(1-s)$. Since the contour integral in (4) is not known in closed form, the numerical evaluations of $P_{\mathrm{dv}}(d_0)$ we shall present in Section VI are based on a recursion-based z-transform inversion [20, Eq. (10)] of $G_{\mathsf{D}}(s)/(1-s)$.

A reduced-complexity approach to compute the delay violation probability from (3) is through the saddlepoint method [21, Eq. (2.2.10)], which, under the assumption that $\lceil d_0/n \rceil > \mathbb{E}[\mathsf{D}] = \lim_{s\uparrow 1} G'_{\mathsf{D}}(s)$, results in the following approximation[2]

$$P_{\mathrm{dv}}(d_0) \approx \frac{B_0(\theta\sigma(\theta))}{\sigma(\theta)(1-e^{-\theta})} e^{\kappa(\theta)-\theta\lceil d_0/n \rceil}. \quad (5)$$

In (5), $\kappa(x) = \log(G_{\mathsf{D}}(e^x))$, $\theta = \arg\min_{x\in\mathbb{R}} \kappa(x) - x\lceil d_0/n \rceil$, $\sigma(x) = \sqrt{\kappa''(x)}$, and $B_0(x) = xe^{x^2/2}Q(x)$, where $Q(x)$ is the Gaussian Q-function and the prime notation denotes derivatives.

We present next, for comparison purposes, an upper bound on (2) obtained through a stochastic-network calculus approach [15]. The proof of this bound, which is easier to evaluate than (4) but less tight, involves specializing the general result reported in [5, Thm. 1] to our setup.

Theorem 2: The delay violation probability $P_{\mathrm{dv}}(d_0)$ in (2) is upper-bounded as

$$P_{\mathrm{dv}}(d_0) \leq \inf_{\substack{s>1:\\ G_{\mathsf{B}}(s)G_{\mathsf{H}}(1/s)<1}} \frac{G_{\mathsf{H}}(1/s)^{d-1}}{1-G_{\mathsf{B}}(s)G_{\mathsf{H}}(1/s)}, \quad (6)$$

where $d = \lceil d_0/n \rceil$ and the PGFs $G_{\mathsf{B}}(s)$ and $G_{\mathsf{H}}(s)$ are

$$G_{\mathsf{B}}(s) = (1-\lambda+\lambda s)^n, \quad G_{\mathsf{H}}(s) = \epsilon + (1-\epsilon)s. \quad (7)$$

Proof: By following the analysis in [12, Sec. 4.4], we have

$$P_{\mathrm{dv}}(d_0) \leq \lim_{t\to\infty} \inf_{s>1} \sum\nolimits_{u=0}^{t} G_{\mathsf{B}}(s)^{t-u} G_{\mathsf{H}}(1/s)^{t+d-1-u} \quad (8)$$

$$\leq \inf_{s>1} G_{\mathsf{H}}(1/s)^{d-1} \sum\nolimits_{u=0}^{\infty} G_{\mathsf{B}}(s)^u G_{\mathsf{H}}(1/s)^u. \quad (9)$$

Here, (8) follows from [12, Eqs. (21)–22)]. We obtain (6) by computing the geometric series in (9). ∎

IV. Asynchronous Model and Analysis

We consider a variation of the setup described in Section II, in which, if the buffer is empty when a packet arrives, the corresponding codeword is transmitted starting from the next available CU. We refer to this model as being frame asynchronous. The rationale for this terminology is that, in this setup, frame synchronization between transmitter and receiver needs to be reacquired whenever the buffer is empty.

[2]See [21, p. 27] for an extension to the case $\lceil d_0/n \rceil < \mathbb{E}[\mathsf{D}]$.

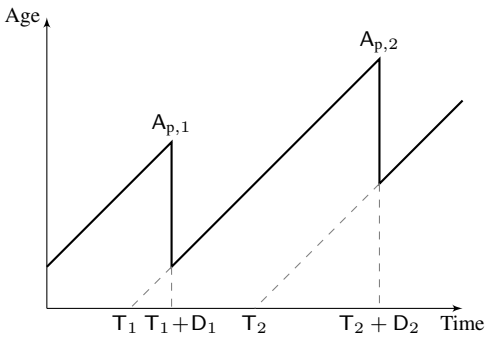

Fig. 1. Peak age of information for the frame-synchronous model: T_m is the frame index corresponding to the arrival of bulk m; $\mathsf{T}_m + \mathsf{D}_m$ is the frame index corresponding to its departure; the peak age $\mathsf{A}_{\mathrm{p},m}$ is the age of information just before the mth bulk departs.

Since in the frame-asynchronous case packets are not grouped into bulks, this setup can be modeled as a $Geo/G/1$ queue. The PGF of the steady-state delay D measured in CUs is given in the following theorem, whose proof follows along the same lines as the proof of Theorem 1.

Theorem 3: For every $\epsilon > 0$ such that $\lambda n < 1-\epsilon$, the PGF of the steady-state delay for the frame-asynchronous model is

$$G_{\mathsf{D}}(s) = \frac{(s-1)(1-\epsilon-\lambda n)s^n}{s-(1-\lambda)-(\lambda+\epsilon(s-1))s^n}. \quad (10)$$

The delay violation probability and its saddlepoint approximation can be obtained by proceeding as in Section III. However, differently from Section III, we cannot obtain a stochastic-network calculus upper bound similar to the one in Theorem 2. Indeed, in the frame-asynchronous setup, the independence assumption made in [5, Lem. 4], which is needed in proof of the delay violation probability upper bound [5, Thm. 1], is violated.

V. Steady-State Peak-Age Violation Probability

We next characterize the violation probability of the steady-state peak age for both the frame-synchronous and frame-asynchronous models.

The peak-age of information is the value of the age of information just before an update is received (see Fig. 1). The peak age $\mathsf{A}_{\mathrm{p},m}$ can be written as [17, Eq. (9)]

$$\mathsf{A}_{\mathrm{p},m} = \max\{\mathsf{D}_{m-1}, \mathsf{T}_m - \mathsf{T}_{m-1}\} + \mathsf{S}_m. \quad (11)$$

In words, it is the sum of the service time of the mth bulk S_m and the maximum between the delay D_{m-1} of the $(m-1)$th bulk of packets and the difference $\mathsf{T}_m - \mathsf{T}_{m-1}$ between the frame indices corresponding to the arrival of the mth and the $(m-1)$th bulks. The PGF of the steady-state peak age $\mathsf{A}_{\mathrm{p}} = \max(\mathsf{D}, \mathsf{T}_2 - \mathsf{T}_1) + \mathsf{S}_1$ can be derived similarly as in [17, Thm. 9] and is given in the next theorem.

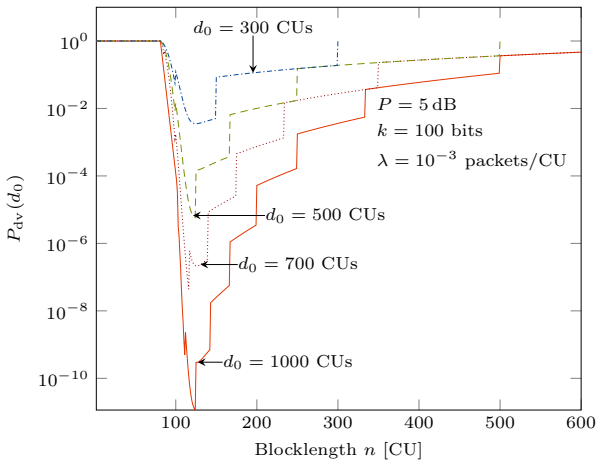

Fig. 2. Delay violation probability vs. blocklength.

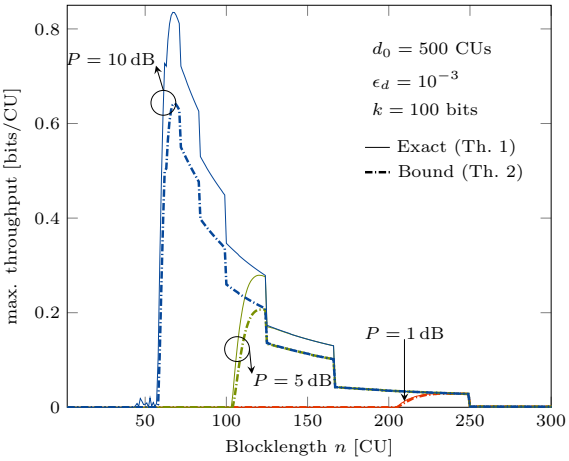

Fig. 4. Maximum throughput vs blocklength for three different SNR values.

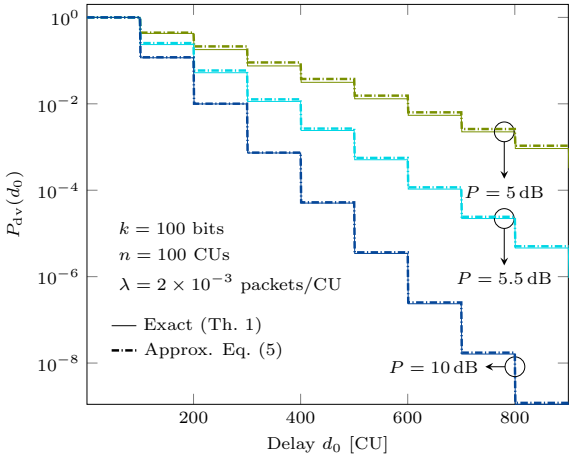

Fig. 3. Delay violation probability and its saddlepoint approximation.

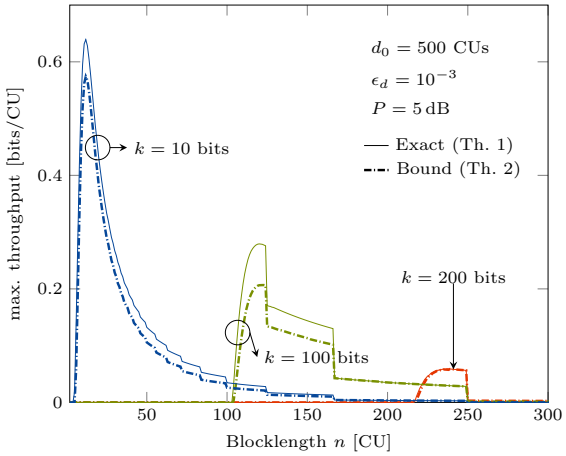

Fig. 5. Maximum throughput versus blocklength for three different values of the information payload k.

Theorem 4: For every $\epsilon > 0$ such that $\lambda n < 1 - \epsilon$, the PGF of the peak age of information $\mathsf{A_p}$ at steady state for the frame-synchronous model is

$$G_{\mathsf{A_p}}(s) = \frac{(1-\lambda+(\lambda-\epsilon)s)^n - (1-\epsilon s)^n(1-\lambda)^n}{(1-\epsilon s)^n(1-(1-\lambda)^n)} \cdot \left(G_{\mathsf{D}}(s) - \frac{(1-s)G_{\mathsf{D}}((1-\lambda)^n s)}{1-(1-\lambda)^n s} \right) \tag{12}$$

where $G_{\mathsf{D}}(s)$ is given in (3). For the frame-asynchronous model the steady state peak age of information is given as

$$G_{\mathsf{A_p}}(s) = \frac{(1-\epsilon)s^n}{1-\epsilon s^n} \left(G_{\mathsf{D}}(s) - \frac{(1-s)G_{\mathsf{D}}((1-\lambda)s)}{1-(1-\lambda)s} \right) \tag{13}$$

where $G_{\mathsf{D}}(s)$ is given in (10).

The peak-age violation probability and the corresponding saddlepoint approximation for both cases can be obtained by proceeding as in Section III.

VI. Numerical Results

Throughout this section, for a given size k (measured in bits) of the information packet and for a given blocklength n, we use Shannon's achievability bound [22, Eq. (3)], which is the tightest achievability bound known for the AWGN channel, to determine the packet error probability $\epsilon(k, n)$. For the parameter range considered in this section, using instead the easier-to-compute normal approximation [1, Eq. (296)] yields very similar results.

In Fig. 2, we illustrate the dependence of the delay violation probability on the blocklength n. We choose $P = 5\,\mathrm{dB}$, $k = 100$ information bits, and arrival rate $\lambda = 10^{-3}$ packets/CU. We observe there exists an optimum blocklength n, and, hence, an optimum code rate k/n, that minimizes the delay violation probability for all values of d_0 considered in the figure. In fact, on the one hand, when the blocklength is small, the packet error probability is large and so is the number of retransmissions, yielding a large probability that the delay exceeds the threshold. On the other hand, when the blocklength is large, the packet error probability is small, but even a small number of retransmissions is sufficient to generate a large delay. Note that the jumps in the plot, which occur at submultiples of d_0, are caused by the change in the number of available retransmission rounds.

A comparison between the delay violation probability (4),

computed through a recursion based z-transform inversion, and the reduced-complexity saddlepoint approximation (5) is drawn in Fig. 3, where it is possible to appreciate that the saddlepoint approximation is extremely accurate.

Next, we study the maximum throughput, which we define as the product $k\lambda^*$ between the number of information bits per packet k and the maximum packet arrival rate λ^* achievable under a constraint on the delay violation probability. In Figs. 4 and 5, the maximum throughput is plotted as a function of the blocklength n for different values of P and k, respectively. In both figures, we set $d_0 = 500$ CUs and a target delay violation probability $\epsilon_d = 10^{-3}$. As a reference, we also plot throughput estimates obtained using the upper bound on the delay violation probability (6), which relies on stochastic network calculus. This bound is accurate only for low SNR or large k. Indeed, for the case $P = 10\,\mathrm{dB}$ and $k = 100$ bits depicted in Fig. 4, the throughput estimate based on (6) is about 20% off. Although the bound (6) provides a loose throughput estimate, it predicts accurately the value of the throughput-maximizing blocklength.

To conclude, we elaborate on the difference between optimizing a system for a target average delay and optimizing it for a target delay violation probability. To this end, we fix $\lambda = 10^{-3}$, $P = 5\,\mathrm{dB}$, $k = 100$ bits, and $d_0 = 500$ CUs. For these parameters, the blocklength values of $n = 100$ CUs and $n = 140$ CUs result in very similar average delays, namely about 154 CUs and 152 CUs, respectively. However, they yield significantly different delay violation probabilities, namely about 1.4×10^{-2} and 2×10^{-4}, respectively. This highlights the importance of performing delay violation probability analyses in latency-critical wireless systems.

Appendix A
Proof of Theorem 1

Let us denote by R_m the number of bulks remaining in the buffer just after the mth bulk leaves the buffer, i.e.,

$$\mathsf{R}_m = \mathsf{Q}_{\mathsf{T}_m + \mathsf{D}_m}. \tag{14}$$

Since the number of bulks arriving in the interval $(\mathsf{T}_m + \mathsf{D}_m, \mathsf{T}_{m+1} + \mathsf{D}_{m+1})$ is independent of R_m, we conclude that $\{\mathsf{R}_m\}_{m=1}^{\infty}$, is a Markov chain governed by

$$\mathsf{R}_{m+1} = \max\{\mathsf{R}_m - 1, 0\} + \sum_{t=1}^{\mathsf{S}_{m+1}} \mathbf{1}\big\{\mathsf{B}_{\mathsf{T}_{m+1}+t} > 0\big\}. \tag{15}$$

Note that the random variables $\{\mathbf{1}\{\mathsf{B}_t > 0\}\}_{t=1}^{\infty}$ and $\{\mathsf{S}_m\}_{m=1}^{\infty}$ are i.i.d., and independent of $\{\mathsf{R}_m\}_{m=1}^{\infty}$. Hence,

$$\mathsf{R}_{m+1} \sim \max\{\mathsf{R}_m - 1, 0\} + \mathsf{U}, \tag{16}$$

where U is the number of bulks of packets arriving during the service time of a bulk, which is given by $\mathsf{U} = \sum_{t=1}^{\mathsf{S}_1} \mathbf{1}\{\mathsf{B}_t > 0\}$. The PGF of the steady-state buffer-size R is [23, Eq. (11.3.11)]

$$G_{\mathsf{R}}(s) = (1 - \mathbb{E}[\mathsf{U}])\frac{(s-1)G_{\mathsf{U}}(s)}{s - G_{\mathsf{U}}(s)},\ \mathbb{E}[\mathsf{U}] < 1. \tag{17}$$

Since $\mathsf{R} \sim \sum_{t=1}^{\mathsf{D}} \mathbf{1}\{\mathsf{B}_t > 0\}$, then $G_{\mathsf{D}}(s) = G_{\mathsf{R}}(G^{-1}_{\mathbf{1}\{\mathsf{B}_1>0\}}(s))$, where

$$G_{\mathbf{1}\{\mathsf{B}_1>0\}}(s) = (1-\lambda)^n + s(1-(1-\lambda)^n). \tag{18}$$

Furthermore, from the definition of U and from (1), we obtain the equality

$$\mathbb{E}[\mathsf{U}] = \mathbb{E}[\mathbf{1}\{\mathsf{B}_1 > 0\}]\,\mathbb{E}[\mathsf{H}_1]\,\mathbb{E}[\mathsf{N}_1] = \lambda n/(1-\epsilon). \tag{19}$$

Next, we observe that $G_{\mathsf{U}}(s) = G_{\mathsf{S}_1}\big(G_{\mathbf{1}\{\mathsf{B}_1>0\}}(s)\big)$, $G_{\mathsf{H}_1}(s) = (1-\epsilon)s/(1-\epsilon s)$, and $G_{\mathsf{S}_1}(s) = G_{\mathsf{N}_1}(G_{\mathsf{H}_1}(s))$ where

$$G_{\mathsf{N}_1}(s) = \frac{(1-\lambda+\lambda s)^n - (1-\lambda)^n}{1-(1-\lambda)^n}. \tag{20}$$

Algebraic manipulations yield (3).

References

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, Apr. 2016.

[3] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[4] E. M. Yeh, "Fundamental performance limits in cross-layer wireless optimization: throughput, delay, and energy," in *Foundations and Trends Commun. Inf. Theory*. Delft, The Netherlands: Now Publisher, 2012, vol. 9, no. 1.

[5] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[6] M. Zafer and E. Modiano, "Minimum energy transmission over a wireless channel with deadline and power constraints," *IEEE Trans. Autom. Control*, vol. 54, no. 12, pp. 2841–2852, Dec. 2009.

[7] R. Singh and P. R. Kumar, "Decentralized throughput maximizing policies for deadline-constrained wireless networks," in *IEEE Annual Conf. Deci., Contr (CDC)*, Osaka, Japan, Dec. 2015, pp. 3759–3766.

[8] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[9] F. Hamidi-Sepehr, H. D. Pfister, and J.-F. Chamberland, "Delay-sensitive communication over fading channels: Queueing behavior and code parameter selection," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 3957–3970, Sep. 2015.

[10] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[11] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wireless Commun. Netw.*, vol. 2013, no. 1, p. 290, Dec. 2013.

[12] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. ACM MSWiM*, Mar. 2015, pp. 13–22.

[13] C. She, C. Yang, and T. Q. Quek, "Cross-layer transmission design for tactile internet," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Washington D.C., U.S.A, Dec. 2016.

[14] 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.

[15] Y. Jiang and Y. Liu, *Stochastic network calculus*. New York, NY, U.S.A.: Springer, 2008.

[16] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[17] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "The stationary distribution of the age of information in FCFS single-server queues," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 571–575.

[18] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, Aug. 2017.

[19] S. K. Bose, *An introduction to queueing systems*. New York, NY, U.S.A.: Springer, 2013.

[20] L. Jenkins, "A useful recursive form for obtaining inverse z-transforms," *Proc. IEEE*, vol. 55, no. 4, pp. 574–575, Apr. 1967.

[21] J. L. Jensen, *Saddlepoint approximations*. Oxford, U.K.: Oxford University Press, 1995.

[22] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, May 1959.

[23] G. Grimmett and D. Stirzaker, *Probability and random processes*. New York, NY, U.S.A.: Oxford University Press, May 2001.