King's Research Portal

Link to publication record in King's Research Portal

Reliable Transmission of Short Packets through Queues and Noisy Channels under Latency and Peak-Age Violation Guarantees

Rahul Devassy, *Student Member, IEEE*, Giuseppe Durisi, *Senior Member, IEEE*,
Guido Carlo Ferrante, *Member, IEEE*, Osvaldo Simeone, *Fellow, IEEE*, Elif Uysal, *Senior Member, IEEE*

Abstract—This work investigates the probability that the delay and the peak-age of information exceed a desired threshold in a point-to-point communication system with short information packets. The packets are generated according to a stationary memoryless Bernoulli process, placed in a single-server queue and then transmitted over a wireless channel. A variable-length stop-feedback coding scheme—a general strategy that encompasses simple automatic repetition request (ARQ) and more sophisticated hybrid ARQ techniques as special cases—is used by the transmitter to convey the information packets to the receiver. By leveraging finite-blocklength results, the delay violation and the peak-age violation probabilities are characterized without resorting to approximations based on large-deviation theory as in previous literature. Numerical results illuminate the dependence of delay and peak-age violation probability on system parameters such as the frame size and the undetected error probability, and on the chosen packet-management policy. The guidelines provided by our analysis are particularly useful for the design of low-latency ultra-reliable communication systems.

Index Terms—Ultra-reliable low-latency communications, ARQ, HARQ, delay violation probability, peak-age violation probability.

I. INTRODUCTION

Emerging wireless applications, such as automated transportation, industrial automation and control, and tactile internet, require the availability of mission-critical links that are able to deliver short information packets within stringent latency and reliability constraints. The requirements for ultra-reliable low-latency communications (URLLC) defined by the international telecommunication union (ITU) [2] underpin the introduction of the URLLC service in the next-generation wireless cellular system (5G).

Considering the isolated transmission of a single packet, new nonasymptotic tools in information theory have been applied

for the design of URLLC since they can capture constraints on very high reliability, short coded packets, as well as short information payloads and sporadic transmissions. Specifically, as recently shown in [3], [4], finite-blocklength information theory provides accurate tools for describing the trade-offs among latency, reliability, and rate when transmitting individual short packets.

Communication latency is, however, not only determined by the blocklength, but also by the contribution of the queuing delays accrued in the presence of a data stream. Queuing delays are inherently random and designs that only control the average latency are not suitable to capture the stringent performance requirements of mission-critical applications. Instead, solutions for URLLC should ensure that the overall delay is below a tolerable threshold with a sufficiently large probability. Given the critical role of queuing delays in guaranteeing URLLC latency performance, we provide in this paper a joint coding-queuing analysis of the probability that the overall steady-state delay, including both queuing and transmission, exceeds a desired level for a given reliability constraint.

As discussed, the overall latency is a key metric of interest in many use cases of URLLC. Latency is measured at the link layer and provides a useful quality-of-service measure for higher layers of the protocol. However, in some important applications, minimizing the delivery latency may not align with the requirements of the application layer. For example, in factory automation, the information packets exchanged over the wireless medium may carry sensor data needed to track a remote process at a given destination. The primary performance metric at the application layer in such a scenario is not the delay of each packet, but rather the freshness of the sensor data available at the destination. Packets that contain outdated sensor data are not valuable to the destination, and insisting on transmitting them with low latency is generally suboptimal. A more relevant performance metric is the *peak age of information*, which measures the maximum elapsed time since the last received update at the destination (see, e.g., [5] and references therein). To address this scenario, in this paper we also investigate the probability that the steady-state peak age of the information packets exceeds a desired level for a given reliability constraint.

Throughout the paper, we consider the point-to-point communication system with random information-packet arrivals per channel use depicted in Fig. 1. The analysis assumes a single-server queue and the use of general variable-length-stop-feedback (VLSF) codes for information transmission [4,
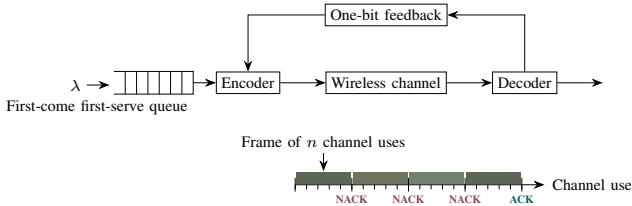
Fig. 1. In the considered system model, a new information packet arrives in each channel use with probability λ. The packet enters a first-come first-serve queue and it is then transmitted over a wireless channel using a variable-length stop-feedback code.

Eq. (10)]. In a VLSF coding scheme, each codeword, which consists of an arbitrary large number of coded symbols, is divided into frames of n symbols. After receiving a frame, the decoder attempts to recover the information packet. Then, it communicates the outcome of the decoding attempt to the encoder through an ACK/NACK bit transmitted over a feedback channel. If decoding fails, the next block is transmitted. If decoding succeeds, transmission is stopped and the information packet is removed from the queue. VLSF codes include as special cases strategies commonly employed in current wireless systems such as simple automatic repetition request (ARQ), where all frames corresponding to the same information packet contain the same coded bits, and incremental-redundancy hybrid ARQ (HARQ), where instead each new frame contains additional parity-check symbols.

Related Work: Aside from the work by Telatar and Gallager [6], who employed an error-exponent approach, most queuing analyses 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 quasi-static 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 Chernoff bound via network calculus or large-deviation theory (see [7], [8] and references therein); and (iii) analyses of the throughput-delay tradeoff under deadline constraints [9], [10]. However, the bit-pipe abstraction is not suitable for the analysis of URLLC traffic, given that 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 [11], and using them may result in inaccurate delay estimates.

Recognizing these limitations, Hamidi-Sepehr *et al.* [12] analyzed the queuing behavior when BCH codes are used. They evaluated both the probability distribution of the steady-state queue size and the average delay. Their analysis accounts for undetected error events, i.e., the event that an ACK is fed back although the decoded message is wrong. Specifically, the authors showed how to mitigate the negative impact of such events by using the erasure decoding rule proposed in [13].

A different approach, which relies on random coding, but is limited to the analysis of ARQ systems, is to replace capacity or outage capacity with the more accurate second-order fixed-blocklength approximations obtained in [3], [14]. This approach has been used in [15] to study the throughput achievable over

a fading channel under a constraint on the probability of buffer overflow; in [16] to analyze the packet delay violation probability in the presence of imperfect channel-state information at the transmitter; and in [17] to design the downlink of an ultra-reliable transmission system under a constraint on the end-to-end delay. All these works rely on the characterization of tail probabilities using Chernoff bound via effective capacity [18], stochastic network calculus [19], and effective bandwidth [20], and provide tight delay estimates only for large delay. Furthermore, these results rely on the simplifying assumption that errors are perfectly detected at the decoder, i.e., the undetected error probability is zero.

An effective-capacity analysis of general HARQ systems is given in [21]. However, this analysis relies on an asymptotic information-theoretic approach based on the renewal-reward theorem (see [22] for details), which is again not suitable for URLLC, since it requires large packets.

A nonasymptotic random-coding lower bound on the rate achievable with VLSF codes for a given error probability and a given average blocklength has been recently proposed in [4, Thm. 3]. As we shall see, this bound will turn out central to our analysis of general HARQ systems.

As mentioned, for applications in which packets carry status updates, the maximum time elapsed since the latest update available at the destination was generated at the source, commonly referred to as peak age of information, is more relevant than delay. Most previous analyses of the peak age of information rely on simple physical-layer models. A recent exception is [23], where the stationary distribution of the peak age is characterized, and [24], where generalized age penalty functions are analyzed. In these works, only an abstract model for the service process is considered. For the special case of the binary erasure channel, where the undetected error probability is zero, analyses of the age for HARQ systems have been recently reported in [25], [26].

Contributions: In this work, we analyze the delay and the peak-age violation probabilities achievable over a wireless channel in which information packets arrive in each channel use according to a discrete-time stationary memoryless Bernoulli process with parameter λ. We assume that the packets are transmitted using a general VLSF coding scheme including ARQ and incremental-redundancy HARQ as special cases. Our specific contributions are as follows:

- Under first-come-first-serve (FCFS) queuing policy, which is optimal in terms of latency, we provide a novel definition of the steady-state delay violation probability that accounts not only for the event that the queuing plus the service time of a packet exceeds a given latency, but also for the undetected error events. The sum of the complementary cumulative distribution function (CCDF) of the steady-state delay at the desired maximum latency and the undetected error probability is then used as a numerically computable upper bound on the steady-state delay violation probability.

- For a fixed VLSF code, characterized by random stopping time τ, frame size n, and undetected error probability ϵ_{u}, we determine in closed form the probability generating function (PGF) of the steady-state delay as a function of the packet arrival rate λ and of the PGF of the stopping time τ. The steady-state delay CCDF is then computed numerically through an

efficient inverse transform. We also show how to accurately approximate this transform using saddlepoint methods.

- For the case of simple ARQ with perfect error detection, we illustrate how to evaluate the delay violation probability using random-coding bounds from finite-blocklength information theory. Specifically, we use the random-coding union bound with parameter s (RCUs) [27, Thm. 1]. The resulting expression allows us to determine numerically the frame size n that minimizes the delay violation probability for a given packet arrival rate λ. Our analysis reveals that an optimization of the frame size based only on average delay fails to minimize the delay violation probability. In particular, considering a binary-input AWGN channel (bi-AWGN), we show that there exist two frame-size values resulting in the same average delay but yielding delay violation probabilities that differ by two orders of magnitude. Finally, we show that the delay violation probability estimates based on the large deviation tools used in [28] are loose, especially at low SNR.

- For the more general case of HARQ, we adapt the random-coding VLSF achievability bound in [4, Thm. 3], which relies on threshold decoding, to obtain an estimate of both the PGF of the stopping time τ and the corresponding undetected error probability ϵ_{u}. We show how to optimize the decoding threshold to trade optimally between the reduction of the steady-state delay CCDF at the desired maximum latency and the reduction of the undetected error probability, so as to minimize the overall delay violation probability.

- We extend the analysis above to the scenario where one is interested in minimizing the peak-age violation probability, rather than the delay violation probability. Focusing on the case of ARQ with perfect packet error detection, we study the impact of the three different queue management policies considered in [5] on the peak age of information violation probability. These policies are: (i) FCFS with system capacity 1 (FCFS-1), where information packets arriving while a packet is served are discarded; (ii) FCFS with system capacity 2 (FCFS-2), where the system contains only the packet that is currently served and the first one arrived afterwards; and (iii) last-come first-serve with preemption in queue (LCFS-Q), where the system contains only the packet that is currently served and the last arrived packet. Our analysis supports the observation reported in [5] that the LCFS-Q policy outperforms the other two. Additionally, we also analyze a policy that is more suited to URLLC, in which whenever a new packet arrives, the transmission of the current packet is interrupted and the transmission of the new packet is started in the next frame. This policy, which we refer to as LCFS with preemption in service (LCFS-S), turns out to outperform the other three policies in terms of peak age violation probability.

The analysis in this paper extends the one reported in [1], which was focused only on ARQ with perfect error detection and queues with FCFS discipline.

Notation: Uppercase sans-serif 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 ceiling 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, $\mathrm{Binom}(n,p)$ 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 $G_{\mathsf{X}}(s) = \mathbb{E}[s^{\mathsf{X}}]$. We use x^n to denote the vector $(x_1, x_2, \ldots, x_n)$ where $n \geq 1$; finally, for a real number y, we let $[y]^+ = \max(0,y)$.

II. SYSTEM MODEL

We consider a general point-to-point discrete-time stationary memoryless channel with probability law[1] $P_{\mathsf{Y}|\mathsf{X}}$. Assume that a new information packet arrives in a given channel use with probability λ and that packet arrivals are independent across channel uses. As a result, the average interarrival time between information packets is $1/\lambda$. Upon its arrival, each packet, which is assumed to carry k information bits, is stored at the transmitter in a single-server queue operating according to the FCFS policy.[2]

An information packet that is ready to be transmitted over the channel is mapped into a coded packet by means of a VLSF encoder. Specifically, each of the 2^k possible messages carried by an information packet is assigned to a distinct codeword of infinite length. Then, the first frame, consisting of the initial n symbols of the codeword associated to the desired message, is transmitted. The VLSF decoder uses a stopping rule to decide whether the corresponding n received symbols are sufficient to decode the message. If they are not sufficient, the decoder sends a NACK message to the encoder via the feedback link, after which the encoder sends the next frame of n symbols of the codeword. This procedure continues until the stopping rule is triggered and the decoder produces an estimate of the transmitted message. Finally, the decoder sends an ACK message to the encoder, which removes the packet from the queue and transmits the next packet in the queue.

We shall denote by τ the random number of frames needed for the transmission of a packet according to the described VLSF scheme. Hence, the packet service time measured in channel uses is $n\tau$. Importantly, a packet may not necessarily be delivered correctly at the end of the service time, i.e., when an ACK is issued. This occurs whenever the decoding process results in a wrong message estimate. We denote the probability of this event, which corresponds to an undetected error, as ϵ_{u}. Throughout, a VLSF coding scheme described as above is referred to as a $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code where P_τ is the probability distribution of τ.

We start by considering a *frame-synchronous* system, where time is organized into time frames of n channel uses and the transmission of a codeword starts only at the beginning of a frame. Under this assumption, if an information packet arrives when the buffer is empty, its transmission is delayed to the next available frame. In Section IV, we shall then relax this assumption and allow transmission to start in the next available channel use when the buffer is empty. We refer to this alternative setup as *frame asynchronous*. This alternative model yields a reduction in latency at the cost of a more involved frame-synchronization procedure. In this paper, we shall assume synchronization to be ideal.

---

[1]Extensions of the system model to channels with memory, which are relevant in the presence of fading, are discussed in Section VII-A.

[2]We will consider more general packet-management policies in Section V.

Under the frame-synchronous assumption, the system can be modeled as a $Geo/G/1$ queue with bulk arrivals, which is sometimes denoted $Geo^{[X]}/G/1$ (see [29, Sec. 4.6.2]). This queue evolves along the time index t running over the time frames. We next elaborate on this by detailing the arrival and departure processes. All packets arriving within a time frame constitute a *bulk*. Let B_t be the number of packets received in the t-th time frame. It follows that the bulk-arrival process $\{\mathsf{B}_t\}_{t=1}^\infty$ is stationary memoryless with $\mathrm{Binom}(n, \lambda)$ marginal distribution. When $\mathsf{B}_t > 0$, we say that a bulk has arrived at time frame t. The arrival time T_m of the mth bulk is, hence, defined as $\mathsf{T}_m = \min\{u : u > 0, \sum_{t=1}^u \mathbf{1}\{\mathsf{B}_t > 0\} = m\}$. The number of packets in the mth bulk is $\mathsf{N}_m = \mathsf{B}_{\mathsf{T}_m}$. 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. The service time of the mth bulk, i.e., the total number of frames needed to transmit all packets in the mth bulk, is denoted by S_m. The service process $\{\mathsf{S}_m\}_{m=1}^\infty$ is stationary memoryless 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. and independent of N_m, and each H_k is the random variable representing the number of time frames needed to transmit one packet. Since each packet is served using the same $(n, k, P_\tau, \epsilon_\mathrm{u})$–VLSF code, we have that H_k has probability distribution P_τ.

We denote by Q_t the number of bulks remaining in queue at the start of the $(t + 1)$th time frame. Finally, the delay $\mathsf{D}_m = \mathsf{W}_m + \mathsf{S}_m$ of the mth bulk (measured in frames) is the sum of the waiting time W_m and the service time S_m. For this queuing system, the process $\{\mathsf{D}_m\}_{m=1}^\infty$ admits a steady-state distribution as long as the *traffic intensity* $\lambda n \, \mathbb{E}[\tau]$, i.e., the ratio between average service time and average packet interarrival time, satisfies $\lambda n \, \mathbb{E}[\tau] < 1$ [30, Thm. 11.3.5]. Indeed, if $\lambda n \, \mathbb{E}[\tau] < 1$, the Markov process describing the queue size is ergodic and, hence, it admits a stationary distribution. This implies that $\{\mathsf{D}_m\}_{m=1}^\infty$ has a steady-state distribution, which is studied in the next section. The frame-asynchronous setup will be described in Section IV.

III. STEADY-STATE DELAY VIOLATION PROBABILITY

In this section, we shall define the steady-state delay violation probability and provide methods to compute it. As detailed below, our definition accounts for both the probability that the delay of a packet exceeds a given latency constraint and for undetected error events. We start by formally defining an $(n, k, P_\tau, \epsilon_\mathrm{u})$–VLSF code. Then, we provide an upper bound on the steady-state delay violation probability that depends on the average packet interarrival rate λ, on the PGF of the decoding time τ, and on the undetected error probability ϵ_u of the underlying VLSF code. Finally, we relate ϵ_u to the PGF of τ by deriving a variable-length random-coding bound based on the threshold decoding scheme proposed in [4, Thm. 3]. We shall also consider the special case of ARQ with perfect error detection, i.e., with $\epsilon_\mathrm{u} = 0$. For this case, we characterize the steady-state delay violation probability by leveraging the fixed-length nonasymptotic information-theoretic bound on the smallest error probability achievable for a given

frame length n and a given number of information bits k provided in [27, Thm. 1].

A. Definition of a VLSF Code

The formal definition of a VLSF code provided next is similar to the one given in [4], with the difference that, in our setup, decoding is attempted after each frame of n channel uses, whereas in [4] decoding is attempted at every channel use. Furthermore, we parametrize the code by using the probability distribution P_τ of the random stopping time τ, rather than its average $\mathbb{E}[\tau]$.

Definition 1: An $(n, k, P_\tau, \epsilon_\mathrm{u})$–VLSF code consists of:

1) A random variable U, defined on a set $\mathcal{U}$,[3] whose realization is revealed to the encoder and the decoder before the start of transmission. The random variable U acts as common randomness and enables the use of randomized encoding and decoding strategies.

2) A randomized encoder $f : \mathcal{U} \times \mathcal{W} \to \mathcal{X}^\infty$, where $\mathcal{W} = \{1, \ldots, 2^k\}$, which maps the message $\mathsf{M} \in \mathcal{W}$ and the common randomness U to an infinite-length codeword with symbols belonging to the set $\mathcal{X}$. We assume that M is uniformly distributed on the set $\mathcal{W}$.

3) A sequence of decoders $g_t : \mathcal{U} \times \mathcal{Y}^{nt} \to \mathcal{W}$, $t \geq 1$ that provide the estimate $\widehat{\mathsf{M}}$ of the message M at the tth decoding attempt, on the basis of the nt dimensional received signal vector $\mathsf{Y}^{nt} \in \mathcal{Y}^{nt}$.

4) A nonnegative integer-valued random variable τ, which is a stopping time that depends only on the common randomness U and on the received frames.

5) The final estimate of M

$$\widehat{\mathsf{M}} = g_\tau(\mathsf{U}, \mathsf{Y}^{n\tau}) \tag{2}$$

which satisfies the error-probability constraint $\Pr\{\widehat{\mathsf{M}} \neq \mathsf{M}\} \leq \epsilon_\mathrm{u}$.

B. Definition of the Delay Violation Probability

For the practically relevant scenario considered in this paper in which the VLSF decoding process may yield undetected error events, it is not satisfactory to define the delay violation probability simply as the CCDF of the steady-state delay—the sum of service time τ and queuing time of a packet at steady state. This is because an undetected error event results in a packet being removed from the queue, although it has not been correctly delivered. To overcome this problem, we define the delay violation probability so as to account for both the event of the delay exceeding a maximum latency and the case of the message being decoded incorrectly.

Formally, for a given latency constraint d_0 measured in channel uses, we define the steady-state delay violation probability $P_\mathrm{dv}(d_0)$ for the case $\lambda n \, \mathbb{E}[\tau] < 1$ as

$$P_\mathrm{dv}(d_0)$$
$$= \limsup_{m \to \infty} \left[1 - \Pr\left\{ \mathsf{D}_m^{(\mathrm{p})} < d_0/n \text{ and } \mathsf{M}_m = \widehat{\mathsf{M}}_m \right\} \right], \tag{3}$$

---

[3]We will discuss the cardinality of this set at the end of Section III-D.

where $\mathsf{D}_m^{(\mathrm{p})}$ denotes the delay of the mth packet measured in number of frames, and M_m and $\widehat{\mathsf{M}}_m$ are the message corresponding to the mth packet and its estimate (2) at the decoder, respectively.[4]

Analyzing (3) directly seems to be prohibitive, apart from the special case $\epsilon_{\mathrm{u}} = 0$, which will be discussed in Section III-C. Therefore, we will focus in the remainder of the paper on the following upper bound $U_{\mathrm{dv}}(d_0)$ on $P_{\mathrm{dv}}(d_0)$, which is obtained by applying the union bound:

$$P_{\mathrm{dv}}(d_0)$$

$$\leq \limsup_{m \to \infty} \left[\Pr\left\{ \mathsf{D}_m^{(\mathrm{p})} \geq d_0/n \right\} + \Pr\left\{ \mathsf{M}_m \neq \widehat{\mathsf{M}}_m \right\} \right] \quad (4)$$

$$\leq \limsup_{m \to \infty} \Pr\{ \mathsf{D}_m \geq d_0/n \} + \epsilon_{\mathrm{u}} \quad (5)$$

$$= \Pr\{ \mathsf{D} \geq d_0/n \} + \epsilon_{\mathrm{u}} \quad (6)$$

$$\triangleq U_{\mathrm{dv}}(d_0). \quad (7)$$

Here, (5) follows because the delay of a bulk coincides with the delay of the last packet in the bulk; and in (6), D denotes a random variable whose distribution coincides with the steady-state distribution of D_m. The upper bound $U_{\mathrm{dv}}(d_0)$ is the sum of the CCDF of the steady-state delay, computed at the latency constraint, and the undetected error probability. In the reminder of this section, we will discuss the computation of these two quantities.

Before doing so, however, it is appropriate to point out that other notions of delay violation probability are possible. For example, if each message carries mission-critical information, it may be preferable to avoid decoding the message, hence violating the latency constraint, rather than decoding it incorrectly. In such a scenario, one may replace the sum in (6) by another suitably chosen function (e.g., a weighted sum). Our analysis can be readily extended to such a scenario.

We start by evaluating the CCDF of the steady-state delay. As a first step, we provide in Theorem 1 below the PGF of D as a function of the packet arrival rate λ and of the PGF of the service time τ.

Theorem 1: For a given packet interarrival rate λ and for every $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code satisfying the stability condition $\lambda n \, \mathbb{E}[\tau] < 1$, the PGF of the steady-state delay D for the frame-synchronous model is

$$G_{\mathsf{D}}(s) = (1 - \lambda n \, \mathbb{E}[\tau])$$
$$\cdot \frac{(1-s)((1-\lambda)^n - (1-\lambda+\lambda G_\tau(s))^n)}{(1-(1-\lambda)^n)(s - (1-\lambda+\lambda G_\tau(s))^n)}. \quad (8)$$

Proof: See Appendix A. ∎

The CCDF of D can be obtained from the PGF in (8) using the inversion formula

$$\Pr\{ \mathsf{D} \geq d_0/n \} = 1 - \left(\frac{\mathbf{1}\{d \geq 2\}}{2\pi i} \oint_{\mathcal{C}} \frac{G_{\mathsf{D}}(s)}{(1-s)s^{d-1}} ds \right), \quad (9)$$

where $d = \lceil d_0/n \rceil$ and $\mathcal{C}$ is a circle centered at the origin enclosing all poles of $G_{\mathsf{D}}(s)/(1-s)$. Since the contour integral in (9) is not known in closed form, the numerical evaluations of (8) we shall present in Section VI are based on a recursion

based z-transform inversion of $G_{\mathsf{D}}(s)/(1-s)$. This method, which is based on [31, Eq. (10)], is detailed in Appendix B.

A reduced-complexity approach to compute the delay violation probability from (8) is through the saddlepoint method [32, Eq. (2.2.10)]. Let $\mathbb{E}[\mathsf{D}]$ be the expectation of the steady-state delay. Note that we can obtain $\mathbb{E}[\mathsf{D}]$ from the PGF of D in (8) by computing the limit $\mathbb{E}[\mathsf{D}] = \lim_{s \uparrow 1} G_{\mathsf{D}}'(s)$ where we use the prime notation to denote derivatives. To provide the saddlepoint approximation, we distinguish between the two cases $\lceil d_0/n \rceil > \mathbb{E}[\mathsf{D}]$ and $\lceil d_0/n \rceil < \mathbb{E}[\mathsf{D}]$. In the first case, we have

$$\Pr\{\mathsf{D} \geq d_0/n\} \approx \frac{B_0(\theta\sigma(\theta))}{\sigma(\theta)(1 - e^{-\theta})} e^{\kappa(\theta) - \theta \lceil d_0/n \rceil}, \quad (10)$$

where $\kappa(x) = \log(G_{\mathsf{D}}(e^x))$, $\theta = \arg\min_{x \in \mathbb{R}} \kappa(x) - x \lceil d_0/n \rceil$, $\sigma(x) = \sqrt{\kappa''(x)}$, and $B_0(x) = x e^{x^2/2} Q(x)$, where $Q(x)$ is the Gaussian Q-function. In the second case, we use that $\Pr\{\mathsf{D} \geq d_0/n\} = 1 - \Pr\{-\mathsf{D} \geq -1 - \lceil d_0/n \rceil\}$ and then apply (10) to the second term. As we shall show in Section VI, the approximation (10) is very accurate over a large range of system parameters.

To evaluate $U_{\mathrm{dv}}(d_0)$ using (9) or (10), one needs the PGF of the decoding time τ and the undetected error probability ϵ_{u}. Next, we discuss how to obtain these quantities. We will start from the simpler case of ARQ with perfect error detection in Section III-C, and then move to the more general VLSF setup with positive undetected error probability in Section III-D.

C. A Special Case: ARQ with Perfect Error Detection

In the ARQ setup, the VLSF encoder repeats the same n channel inputs in each frame until it receives an ACK. We will analyze this coding scheme under the commonly used simplifying assumption of perfect error detection, namely, an ACK is fed back only if the message is decoded correctly. This can be achieved, for example, through the use of hashing methods, whose overhead we shall neglect. We will consider the more general and practically relevant case of VLSF codes with positive undetected error probability in Section III-D. Under the assumption of perfect error detection, we have $\epsilon_{\mathrm{u}} = 0$ and, hence, $U_{\mathrm{dv}}(d_0) = P_{\mathrm{dv}}(d_0)$.

As we shall see shortly, the PGF of τ is uniquely determined in this case by the frame error probability ϵ of the fixed-blocklength (k, n) channel code that is used to map the information packet into the n coded symbols transmitted within a frame. The RCUs bound [27, Thm. 1] provides the following random-coding bound on ϵ, as a function of the code parameters k and n.

Theorem 2 ([27, Thm. 1]): Let X^n be an i.i.d. n-dimensional vector with i.i.d. entries distributed according to P_{X}. For a given $\alpha > 0$,[5] let the generalized information density be defined as

$$\imath_\alpha(x_1^n; y_1^n) = \sum_{k=1}^n \log \frac{P_{\mathsf{Y}|\mathsf{X}}(y_k \,|\, x_k)^\alpha}{\mathbb{E}\left[P_{\mathsf{Y}|\mathsf{X}}(y_k \,|\, \mathsf{X}_k)^\alpha \right]}. \quad (11)$$

Then, there exists a (k, n)-code with average frame error probability ϵ satisfying

$$\epsilon \leq \epsilon_{\mathrm{rcus}} \triangleq \inf_{\alpha \geq 0} \mathbb{E}\left[\exp\left\{ -\left[\imath_\alpha(\mathsf{X}^n; \mathsf{Y}^n) - \log(2^k - 1) \right]^+ \right\} \right]$$
$$(12)$$

---

[4] Note that we define the delay violation event in terms of packet delay rather than bulk delay. This explains the presence of the superscript (p) to the delay random variable $\mathsf{D}_m^{(\mathrm{p})}$. This definition has the advantage to hold for both the frame-synchronous and the frame-asynchronous cases.

[5] This parameter is traditionally denoted as s, hence the name of the bound. We denote it by α since we use s to denote the argument of PGFs.

where the expectation is with respect to the joint probability distribution

$$P_{\mathsf{X}^n\mathsf{Y}^n}(x^n, y^n) = \prod_{k=1}^{n} P_{\mathsf{X}}(x_k)P_{\mathsf{Y}\mid\mathsf{X}}(y_k \mid x_k). \quad (13)$$

Using the achievable error probability ϵ_{rcus}, we conclude that the decoding time is distributed as $\tau \sim \mathrm{Geom}(1 - \epsilon_{\mathrm{rcus}})$. Hence, its PGF is $G_\tau(s) = (1 - \epsilon_{\mathrm{rcus}})s/(1 - \epsilon_{\mathrm{rcus}}s)$.

An Alternative Bound Through a Large-Deviation Analysis: A common approach in the queuing literature to characterize the CCDF of the steady-state delay is through the Chernoff bound. When the packet service time is deterministic and constant, this approach is the effective bandwidth method [33], [20], whereas when the arrival rate is deterministic and constant and the service rate is random one uses the effective capacity method [18]. In our model, both the arrival and the service processes are random. Hence, these two tools cannot be used. A suitable tool for our setup, under the additional assumption of ARQ with perfect error detection, is stochastic network calculus [19]. In the next theorem, we present an upper bound on $P_{\mathrm{dv}}(d_0)$ based on stochastic network calculus, which is obtained by adapting [28, Thm. 1] to our setup.[6] As we shall see, this bound is sometimes much looser than the exact characterization of $P_{\mathrm{dv}}(d_0)$ obtainable through (9) or (10).

Theorem 3: For the ARQ setup and under the assumption of perfect error detection, the steady-state delay violation probability $P_{\mathrm{dv}}(d_0)$ is upper-bounded as

$$P_{\mathrm{dv}}(d_0) \leq \inf_{\substack{s>1:\\ G_{\mathsf{B}}(s)G_{\mathsf{H}}(1/s)<1}} \frac{G_{\mathsf{H}}(1/s)^{d-1}}{1 - G_{\mathsf{B}}(s)G_{\mathsf{H}}(1/s)}, \quad (14)$$

where $d = \lceil d_0/n \rceil$ and the PGFs $G_{\mathsf{B}}(s)$ and $G_{\mathsf{H}}(s)$ are $G_{\mathsf{B}}(s) = (1 - \lambda + \lambda s)^n$ and $G_{\mathsf{H}}(s) = \epsilon_{\mathrm{rcus}} + (1 - \epsilon_{\mathrm{rcus}})s$.

Note that the numerical evaluation of (14) has the same complexity as the evaluation of the saddle-point approximation (10), which, as we shall see, is more accurate. Both expressions are easier to evaluate than (9).

D. VLSF Codes with Positive Undetected Error Probability

We present next random-coding characterizations of the CCDF of the stopping time τ of a $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code, and of the corresponding undetected error probability ϵ_{u}, which are based on an adaptation of the threshold-decoding achievability bound provided in [4, Thm. 3].

Theorem 4: Fix a positive real number γ. Let $\mathsf{X} = (\mathsf{X}_1, \mathsf{X}_2, \cdots)$ be a stationary memoryless stochastic process with $\mathsf{X}_1 \sim P_{\mathsf{X}}$. Let $\overline{\mathsf{X}}$ be an independent copy of the same process. Define a sequence of information density functions $\imath(x^t; y^t) = \sum_{k=1}^{t} \log\left(\frac{P_{\mathsf{Y}\mid\mathsf{X}}(y_k \mid x_k)}{P_{\mathsf{Y}}(y_k)}\right)$, $t \geq 1$, where $P_{\mathsf{Y}}(\cdot)$ is the output distribution induced by the distribution P_{X} on the input symbols through the channel law $P_{\mathsf{Y}\mid\mathsf{X}}$. Also define the stopping times

$$\tau^* = \inf\{t \geq 1 : \imath(\mathsf{X}^{nt}; \mathsf{Y}^{nt}) \geq \gamma\} \quad (15)$$

$$\overline{\tau}^* = \inf\{t \geq 1 : \imath(\overline{\mathsf{X}}^{nt}; \mathsf{Y}^{nt}) \geq \gamma\}. \quad (16)$$

Then, there exists an $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code with

$$\Pr\{\tau \geq t\} \leq \Pr\{\tau^* \geq t\}, \quad t \geq 0 \quad (17)$$

and

$$\epsilon_{\mathrm{u}} \leq (2^k - 1)\Pr\{\overline{\tau}^* \leq \tau^*\}. \quad (18)$$

Proof: See Appendix C. ∎

In words, we can think of τ^* as the random variable describing the frame index in which the accumulated information density corresponding to the transmitted codeword exceeds the threshold γ. Similarly, $\overline{\tau}^*$ is the frame index in which the accumulated information density corresponding to a codeword different from the transmitted one exceeds γ. Then, as formalized in (17), τ^* stochastically dominates τ. Furthermore, as shown in (18), we have an undetected error if the inequality $\overline{\tau}^* \leq \tau^*$ holds for some incorrect codeword.

In Section VI, we will use Theorem 4 to evaluate the upper bound $U_{\mathrm{dv}}(d_0)$ on the delay violation probability defined in (7). Note that by replacing in (8) the actual stopping time τ of the threshold-based VLSF code in Theorem 4 with τ^*, one obtains an upper bound on $U_{\mathrm{dv}}(d_0)$ because of the stochastic dominance defined by (17).

Handling Truncated Hybrid ARQ: In Theorem 4, we allow for an infinite number of retransmissions, which is not feasible in practice. Our achievability bound can be easily adapted to include the requirement that the number of retransimission is limited—a scenario sometimes referred to as truncated hybrid ARQ. The corresponding bound is given in the following corollary, whose proof, which is omitted, is a simple adaptation of the proof of Theorem 4.

Corollary 1: Let τ^* and $\overline{\tau}^*$ be as defined in (15) and (16), respectively. Then, there exist an $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code with codewords spanning at most n_{t} frames satisfying

$$\Pr\{\tau \geq t\} \leq \Pr\{\min(n_{\mathrm{t}}, \tau^*) \geq t\}, \ 1 \leq t \leq n_{\mathrm{t}} \quad (19)$$

and

$$\epsilon_{\mathrm{u}} \leq (2^k - 1)\Pr\{\overline{\tau}^* \leq \min(n_{\mathrm{t}}, \tau^*)\} + \Pr\{\tau^* \geq n_{\mathrm{t}} + 1\}. \quad (20)$$

The additional term on the right-hand side of (20) compared to (18) captures the case in which a NACK is sent after the transmission of the last frame, which yields a detected error.

As a side remark, we note that, for the case in which the number of frame transmissions does not exceed a maximum number n_{t}, the probability distribution of τ can be described by providing an $(n_{\mathrm{t}} - 1)$-dimensional vector. This implies that the set of pairs $(P_\tau, \epsilon_{\mathrm{u}})$ for which an $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code exists can be mapped into a subset of $\mathbb{R}^{n_{\mathrm{t}}}$. It follows from Definition 1 that this subset is the convex hull of the set of points in $\mathbb{R}^{n_{\mathrm{t}}}$ corresponding to deterministic (i.e., non-randomized) $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF codes. Applying Carathéodory's theorem, we conclude that every $(n, k, P_\tau, \epsilon_{\mathrm{u}})$–VLSF code can be written as a convex combination of $n_{\mathrm{t}} + 1$ deterministic VLSF codes. This implies that the common randomness U in Definition 1 belongs to a set of cardinality no larger than $n_{\mathrm{t}} + 1$.

Note that a stochastic-network calculus upper bound on $U_{\mathrm{dv}}(d_0)$, similar to the one we reported in Theorem 3 for the case of simple ARQ with perfect error detection, is not available in the more general VLSF setup considered here. Indeed, a crucial

---

[6] The proof of this result, which follows closely the analysis in [28, Sec. 4.4], is omitted due to space constraint.

assumption in the proof of Theorem 3 is that the random process describing the number of packets leaving the queue in a given frame is stationary and memoryless. This assumption does not hold in the general VLSF setup.

IV. THE FRAME-ASYNCHRONOUS SETUP

We next consider an alternative setup in which, if the buffer is empty when a packet arrives, the corresponding codeword is transmitted starting from the next available channel use. We refer to this model as frame asynchronous. This setup can be modeled as a simple $Geo/G/1$ queue. The PGF of the steady-state delay D measured in channel uses is given in the following theorem.

Theorem 5: For a given packet interarrival rate λ and for every $(n, k, P_\tau, \epsilon_u)$–VLSF code satisfying the stability condition $\lambda n \, \mathbb{E}[\tau] < 1$, the PGF of the steady-state delay for the frame-asynchronous model is $G_D(s) = (1 - \lambda n \, \mathbb{E}[\tau])(s-1)G_\tau(s^n)/(s - 1 + \lambda(1 - G_\tau(s^n)))$.

Proof: After setting $N_1 = 1$, $B_1 \sim \text{Bern}(\lambda)$, and noting that S_1 has the same distribution as $n\tau$, one uses the same steps as in the proof of Theorem 1. ∎

The CCDF of the steady-state delay and its saddlepoint approximation can be obtained by proceeding as in Section III-B. Also for this setup, a stochastic network calculus bound on the delay violation probability similar to Theorem 3 is not feasible since the random process describing the number of packets leaving the queue at each channel use is not stationary memoryless.

V. STEADY-STATE PEAK-AGE VIOLATION PROBABILITY

We next characterize the violation probability of the steady-state peak age at the application layer. As in [5], we assume that the destination is interested in timely updates about the status of a random process observed at the source. Each of the information packets generated by the source contains a sample of this random process and the time at which the sample was taken. We focus only on the frame-synchronous model, and assume that the time stamp of a packet is the frame index at which the packet enters the queue. So the mth packet carries the time index $T_m^{(p)}$. Since packets arriving in the same frame carry the same time stamp, we shall assume throughout this section that only one of these packets is allowed to enter the queue, whereas the other packets are discarded. In the terminology we introduced in Section II, this corresponds to assuming that each bulk of packets contains a single packet.

Let now $T_m^{(p)} + D_m^{(p)}$ be the time frame at which the mth transmitted packet leaves the queue. It follows that the index of the most recent packet received at the destination at time frame t is $\widehat{m}(t) = \max\left\{ m : T_m^{(p)} + D_m^{(p)} \le t \right\}$ and the corresponding time stamp is $T_{\widehat{m}(t)}^{(p)}$. The age of information is the discrete-time random process $\bar{A}(t) = t - T_{\widehat{m}(t)}^{(p)}$, $t = 1, 2, \ldots$; the peak-age of information A_m is the value of the age of information just before the update contained in the mth transmitted packet is received (see Fig. 2). As shown in the figure, the peak age A_m is the sum of the delay $D_m^{(p)}$ of the mth transmitted packet and of the difference $T_m^{(p)} - T_{m-1}^{(p)}$ between the frame indices corresponding to the arrival of the $(m-1)$th and the mth transmitted packets. Similar
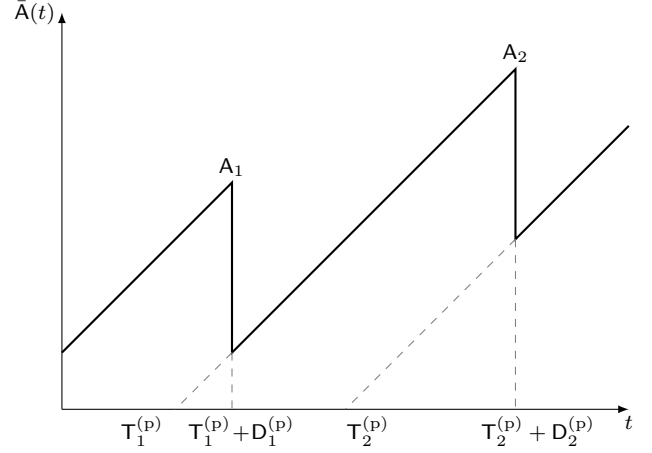

Fig. 2. Peak age of information for the frame-synchronous model: $T_m^{(p)}$ is the frame index corresponding to the arrival of the mth transmitted packet; $T_m^{(p)} + D_m^{(p)}$ is the frame index at which the mth transmitted packet leaves the queue; the peak age A_m is the age of information just before the mth transmitted packet departs.

to (3), we define the steady-state peak-age violation probability as

$$P_{av}(a_0) = \limsup_{m \to \infty} \left[1 - \Pr\left\{ A_m < a_0/n \text{ and } M_m \ne \widehat{M}_m \right\} \right]. \tag{21}$$

Then, we upper-bound it as

$$P_{av}(a_0) \le \Pr\{A \ge a_0/n\} + \epsilon_u \triangleq U_{av}(a_0). \tag{22}$$

Here, A is steady-state peak age of information, defined as $A = D + T_2 - T_1$. Note that we dropped the superscript (p) to keep the notation compact.

Our analysis of the delay violation probability in Section III assumes that the queue operates according to a FCFS discipline. When packets contain status updates and one is interested in the freshness of the information available at the receiver, it is natural to consider other queue management policies that are better suited to minimize the peak-age violation probability.

In what follows, we analyze the steady-state peak-age violation probability for the three packet management policies proposed in [5] in the context of peak-age analyses under continuous-time exponential arrival and service processes: (i) FCFS with system capacity 1 (FCFS-1); (ii) FCFS with system capacity 2 (FCFS-2); and (iii) last come first serve with preemption in queue (LCFS-Q). The FCFS-1 policy discards every newly arrived packet during the transmission of a given packet. Under such a policy, the system can contain at most one packet at any given time. In the FCFS-2 and the LCFS-Q policies, the queue is allowed to hold an additional packet other than the one currently served. In FCFS-2, this additional packet is the first packet arrived in a new frame during the service time of the current packet. In LCFS-Q, it is the last packet arrived in a new frame. These two policies can be modeled by a system with maximum capacity of 2 and nonpreemptive/preemptive buffering, respectively [34]. We also consider a new policy, which we denote as LCFS with preemption in service (LCSF-S). According to this policy, whenever a new packet arrives, the transmission of the current packet is interrupted and the transmission of the new packet is started in the next frame.

The policies FCFS-1 and FCFS-2 involve no preemption; hence, they may be more suitable for energy-constrained applications. LCFS-Q achieves lower peak age violation probability than FCFS-2 because it allows for preemption. However, its queue management is more computationally intensive because a full queue needs to be updated whenever a new packet arrives. Finally, LCSF-S is the most complex queue-management policy, but, as we shall see, it also offers the best performance and appears to be best suited to URLLC.

We shall next derive the PGF of the steady-state peak age for the four packet management policies. The corresponding peak-age violation probability can be obtained by following the steps detailed in Section III. We start with the FCFS-1 policy.

Theorem 6: For a given packet interarrival rate λ and for every $(n, k, P_\tau, \epsilon_u)$–VLSF code satisfying the stability condition $\lambda n \, \mathbb{E}[\tau] < 1$, the PGF of the peak age of information A at steady state with FCFS-1 for the frame-synchronous model is

$$G_{\mathsf{A}}(s) = G_\tau(s)^2 s(1 - (1-\lambda)^n)/((1 - s(1-\lambda)^n)). \quad (23)$$

Proof: Since the packets that arrive during the service time of a given packet are discarded, the interarrival time $\mathsf{T}_2 - \mathsf{T}_1$ is no longer geometrically distributed. It is then convenient to express the peak age as $\mathsf{A} = \mathsf{H}_1 + \mathsf{H}_2 + \mathsf{T}$ where H_1 and H_2 are the service time of two packets and T is the amount of time (measured in number of frames) elapsed from the completion of the service of the first packet until the next packet arrives. These three random variables are independent. Furthermore, $\mathsf{T} \sim \mathrm{Geom}(1 - (1-\lambda)^n)$ in the frame-synchronous case. We obtain (23) by using that the random variables H_1 and H_2 have probability distribution P_τ. ∎

An analysis of the remaining three polices under general VLSF codes appears unfeasible. Hence, we shall focus in the reminder of this section on the simpler case of ARQ with perfect error detection. Under these assumptions, the service time follows a geometric distribution. The memorylessness of the geometric distribution allows us to adapt to our scenario the derivation presented in [5, Secs. IV-B and IV-C] for the continuous-time case under an abstract model for the service time. This is done for FCFS-2 and LCFS-Q in the next theorem.[7]

Theorem 7: Consider the ARQ frame-synchronous model. The PGF of the peak age of information for FCFS-2 and LCFS-Q is

$$G_{\mathsf{A}}(s) = G_{\mathsf{W}}(s)G_{\mathsf{H}}(s)$$
$$\cdot \left(\frac{p_0}{p_0 + p_1} G_{\mathsf{H}^{(0)}}(s)G_{\mathsf{T}}(s) + \frac{p_1}{p_0 + p_1} G_{\mathsf{H}^{(1)}}(s) \right) \quad (24)$$

Here, $\mathsf{T} \sim \mathrm{Geom}(1 - (1-\lambda)^n)$ with PGF

$$G_{\mathsf{T}}(s) = s(1 - (1-\lambda)^n)(1 - (1-\lambda)^n s) \quad (25)$$

is the amount of time elapsed from the completion of the service of a packet until the next packet arrives; W is the waiting time of a packet, whose PGF is given by

$$G_{\mathsf{W}}(s) = \begin{cases} \dfrac{p_0}{p_0 + p_1} + \dfrac{p_1 G_{\mathsf{H}}(s)}{(p_0 + p_1)s}, & \text{for FCFS-2} \\[2ex] p_0 + (1 - p_0)G_{\mathsf{H}^{(0)}}(s)/s, & \text{for LCFS-Q}. \end{cases} \quad (26)$$

---

[7] The proof, which follows from an adaptation to the discrete-time case of the analysis in [5, Secs. IV-B and IV-C] is omitted due to space constraint.
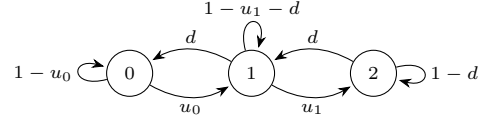

Fig. 3. State transition diagram of the size of the queue for the FCFS-2 and LCFS-Q packet management policies.

In (24) and (26),

$$p_0 = d^2 / \left(d^2 + u_0 d + u_0 u_1 \right) \quad (27)$$

where

$$u_0 = 1 - (1 - \lambda)^n \quad (28)$$
$$u_1 = \epsilon_{\mathrm{rcus}}(1 - (1 - \lambda)^n) \quad (29)$$
$$d = (1 - \epsilon_{\mathrm{rcus}})(1 - \lambda)^n. \quad (30)$$

Furthermore,

$$p_1 = u_0 p_0 / d \quad (31)$$
$$G_{\mathsf{H}}(s) = s(1 - \epsilon_{\mathrm{rcus}})/(1 - \epsilon_{\mathrm{rcus}} s) \quad (32)$$
$$G_{\mathsf{H}^{(0)}}(s) = s(1 - \epsilon_{\mathrm{rcus}}(1 - \lambda)^n)/(1 - \epsilon_{\mathrm{rcus}}(1 - \lambda)^n s) \quad (33)$$
$$G_{\mathsf{H}^{(1)}}(s) = \left((p_0 + p_1)G_{\mathsf{H}}(s) - p_0 G_{\mathsf{H}^{(0)}}(s) \right)/p_1. \quad (34)$$

Some remarks are in order. As shown in [5], in both FCFS-2 and LCFS-Q the queue size can be described by the Markov chain with three states depicted in Fig. 3. The state transition probability u_0 in (28) is the probability that a new packet arrives in a given frame, u_1 in (29) is the probability that a new packet arrives and that a NACK is fed back in a given frame, and d in (30) is the probability that no new packet arrives and an ACK is fed back in a given frame. The parameters p_0 and p_1 in (27) and (31) are the probabilities that the queue has size zero and one in steady state, respectively. Finally, in (32)–(34), $\mathsf{H} \sim P_\tau$ is the service time and $\mathsf{H}^{(0)}$ and $\mathsf{H}^{(1)}$ are the conditional service time given that no new packet and one new packet has entered the queue, respectively.

Finally, we characterize PGF of the peak age of information for LCFS-S in the following theorem.

Theorem 8: Consider the ARQ frame-synchronous model. The PGF of the peak age of information for LCFS-S is

$$G_{\mathsf{A}}(s) = G_{\mathsf{T}^{(0)}}(s) \frac{p_{\mathrm{d}} G_{\mathsf{H}^{(0)}}(s)}{1 - (1 - p_{\mathrm{d}})G_{\mathsf{H}^{(0)}}(s)}. \quad (35)$$

where $G_{\mathsf{H}^{(0)}}(s)$ is defined in (33),

$$p_{\mathrm{d}} = (1 - \epsilon_{\mathrm{rcus}})/(1 - \epsilon_{\mathrm{rcus}}(1 - \lambda)^n) \quad (36)$$
$$G_{\mathsf{T}^{(0)}}(s) = \left(G_{\mathsf{T}}(s) - (1 - p_{\mathrm{d}})G_{\mathsf{H}^{(0)}}(s) \right)/p_{\mathrm{d}}. \quad (37)$$

In (37), the PGF $G_{\mathsf{T}}(s)$ is given in (25).

Proof: See Appendix D. ∎

Some comments on Theorem 8 are in order. The constant p_{d} in (36) is the probability that no preemption in service occurs. The random variable $\mathsf{H}^{(0)}$ is distributed as the conditional service time given that no preemption occurs. The random variable T denotes the number of frames elapsed between two consecutive packet arrivals; finally, the distribution of the random variable $\mathsf{T}^{(0)}$ equals the conditional distribution of T given that no preemption occurs.
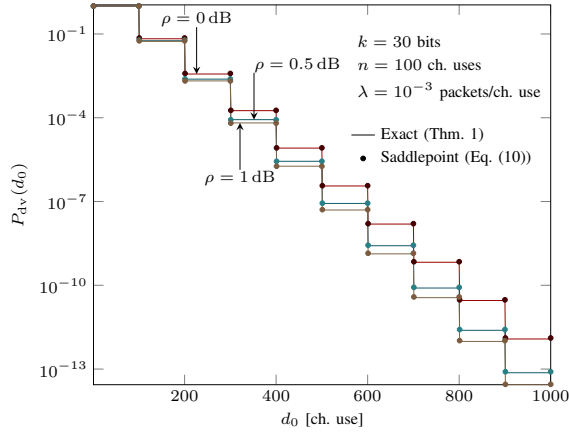
Fig. 4. Delay violation probability for different values of SNR ρ for the case of ARQ with perfect error detection.
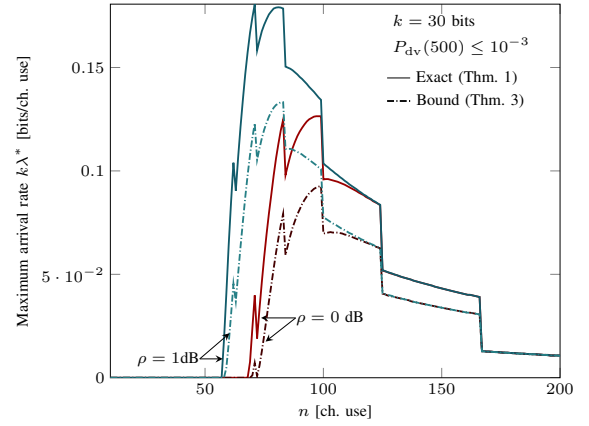

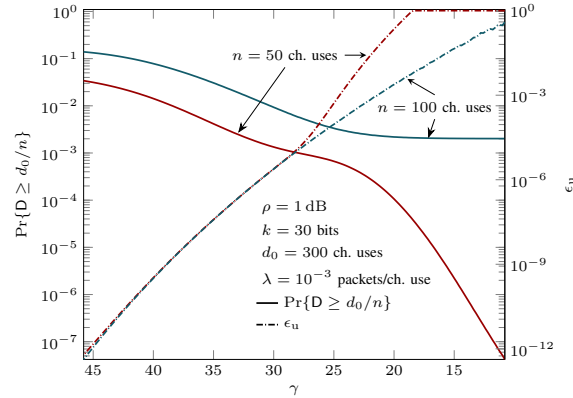Fig. 5. Maximum throughput as a function of the frame size for ARQ with perfect error detection.

Fig. 6. Steady state delay CCDF and undetected error probability for different values of threshold γ for general VLSF setup.

VI. NUMERICAL RESULTS

Throughout this section, we consider a binary-input AWGN channel with input $x \in \{-\sqrt{\rho}, \sqrt{\rho}\}$ for $\rho > 0$, output $y \in \mathbb{R}$ and channel law $P_{\mathsf{Y}|\mathsf{X}}(y \,|\, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}}$. Note that, since the variance of the Gaussian additive noise is one, the parameter ρ defines the signal-to-noise ratio (SNR). This choice allows us to focus on the impact of key parameters such as number of information bits, frame size, etc., without the need to make additional choices such as pilot design for fading channels or antenna processing for multiantenna systems. Nevertheless, it is important to highlight that our analysis holds for more general channel models that include fading, multiple antennas, pilot-assisted transmission, and mismatch decoding as well as interference. Such extensions, which are relevant for URLLC, are discussed in Section VII-A.

We provide next numerical investigations aimed at determining the dependence of the delay violation and the peak-age violation probability on system parameters such as SNR, frame size, undetected error probability as well as packet-management policy for the case of age of information. We will also illustrate the tightness of the saddlepoint approximation (10) and discuss, for ARQ with perfect error detection, the accuracy of the stochastic-network calculus bound given in Theorem 3. Throughout, we will assume that the information packets carry $k = 30$ bits unless specified otherwise. For space constraint, we shall focus uniquely on the frame-synchronous setup.

Simple ARQ with Perfect Error Detection: In Fig. 4, we study the dependence of the delay violation probability $P_{\mathrm{dv}}(d_0)$ on the SNR ρ for the case of ARQ with perfect error detection. We assume a frame size $n = 100$ channel uses and an arrival rate $\lambda = 10^{-3}$ packets/channel use. The delay violation probability is computed using both a recursion-based z transform inversion, which is used to evaluate (9) and the reduced-complexity saddlepoint approximation (10). As one can see from the figure, the saddlepoint approximation is extremely accurate. This turns out to be true also in the more general VLSF setup. When the SNR increases, the channel becomes more reliable and, hence, both the service time and the overall packet delay decrease.

Next, we study the maximum throughput achievable under a constraint on the delay violation probability. The throughput is

defined as the product $k\lambda^*$ between the number of information bits per packet k and the maximum packet arrival rate λ^*. In Fig. 5, we plot the throughput $k\lambda^*$ as a function of the frame size n for different values of ρ. In the figure, we set $d_0 = 500$ channel uses and a target delay violation probability of 10^{-3}. As a reference, we also plot throughput estimates obtained by using the upper bound on the delay violation probability (14), which relies on stochastic network calculus. From the figure we see that, for the case $\rho = 0\,\mathrm{dB}$, the throughput estimate based on (14) is about 35% lower than what predicted by our exact analysis. The figure illustrates that there exists a throughput maximizing frame size. In fact, on the one hand, when the frame size 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 for a given arrival rate. On the other hand, when the frame size is large, the packet error probability is small, but even a small number of retransmissions is sufficient to generate a large delay for a given arrival rate. Note that the discontinuities in the plot, which occur at submultiples of d_0, are caused by the change in the number of available retransmission rounds. Finally, although the bound (14) provides a loose throughput estimate, it seems to predict accurately the value of the throughput-maximizing frame size.

General VLSF Codes: We now consider VLSF codes combined with the threshold decoder used in the proof of Theorem 4.
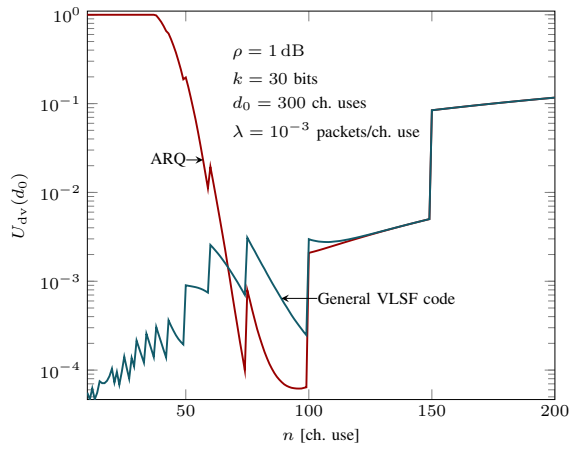
Fig. 7. Delay violation probability versus frame size for both ARQ with perfect error detection and general VLSF setup.

Fig. 8. Peak-age violation probability as a function of the packet arrival rate for different packet-management policies.

Fig. 6 plots the CCDF of the steady-state delay and the undetected packet error probability ϵ_{u} as a function of the decoding threshold γ. We recall that the upper bound $U_{\mathrm{dv}}(d_0)$ on the delay violation probability consists of the sum of these two quantities—see (7). When the threshold γ is low, the probability that the delay exceeds the latency constraint is small, because the decoding process terminates early, but the undetected error probability is high. In fact, it is likely that a codeword different from the transmitted one has an accumulated information density that exceeds the threshold. In contrast, as the threshold γ increases, the undetected error probability decreases but the probability that the delay is above threshold increases.

In Fig. 7, we plot the delay violation probability versus the frame size for both ARQ with perfect error detection and the general VLSF setup. Note that the comparison is not fair because of the perfect-error detection assumption in the ARQ setup. Yet, one sees that for small frame sizes, VLSF codes yield lower delay violation probability. This is because choosing the frame size too small or too big penalizes performance of ARQ, whereas decreasing the frame size enhances the performance of VLSF since it allows the decoder to sample the accumulated information density more frequently.

Finally, 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 consider ARQ with perfect error detection and focus on the parameters in Fig. 7. For these parameters, the frame size values of $n = 60$ channel uses and $n = 91$ channel uses result in very similar average delays, namely about 95.8 and 96 channel uses, respectively. However, they yield significantly different delay violation probabilities, namely about 1.38×10^{-2} and 6.44×10^{-5}, respectively. This highlights the importance of performing delay violation probability analyses in latency-critical wireless systems.

Peak-Age Violation Probability: In Fig. 8, we illustrate the peak-age violation probability for the case of frame-synchronous system and ARQ with perfect error detection, as a function of the packet arrival rate λ, for the packet-management policies analyzed in Section V, namely FCFS-1, FCFS-2, LCFS-Q, and LCFS-S. We consider both $k = 25$ bits and $k = 30$ bits; as expected, the peak age violation probability increases with
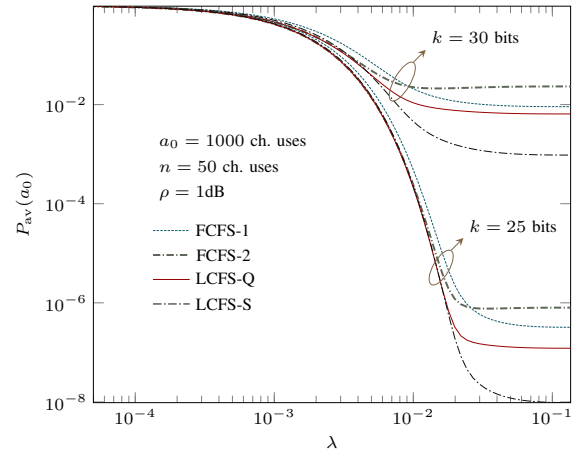
number of information bits. We also observe that LCFS-S is uniformly better than FCFS-1, FCFS-2, and LCFS-Q. Furthermore, FCFS-1 outperforms FCFS-2 at high arrival rates, whereas the situation is reversed at lower arrival rates. Indeed, for high arrival rates, $P_{\mathrm{av}}(a_0)$ converges to

$$\lim_{\lambda \to 1} P_{\mathrm{av}}(a_0)$$
$$= \begin{cases} \Pr\{\mathsf{H}_1 + \mathsf{H}_2 + 1 \geq a_0/n\}, & \text{for FCFS-1} \\ \Pr\{\mathsf{H}_1 + \mathsf{H}_2 + \mathsf{H}_3 - 1 \geq a_0/n\}, & \text{for FCFS-2} \\ \Pr\{\mathsf{H}_1 + \mathsf{H}_2 \geq a_0/n\}, & \text{for LCFS-Q} \\ \Pr\{\mathsf{H}_1 + 1 \geq a_0/n\}, & \text{for LCFS-S} \end{cases}$$

where $\mathsf{H}_1, \mathsf{H}_2$, and H_3 are i.i.d. and $\mathsf{H}_1 \sim \mathrm{Geom}(1 - \epsilon_{\mathrm{rcus}})$. For the case $k = 25$ bits, this limit is 3.2×10^{-7} for FCFS-1, 8×10^{-7} for FCFS-2, 1.2×10^{-7} for LCFS-Q, and 9.7×10^{-9} for LCFS-S. These results support the findings in [5] and extend them to a more realistic physical-layer model as well as to a policy, LCFS-S, that is more suitable to URLLC.

VII. EXTENSIONS OF THE FRAMEWORK

A. Fading and Interference

While the results in the previous section focus on the standard bi-AWGN channel, the analysis can be easily extended to encompass more realistic system models for URLLC transmission, which include fading, multiple antennas, pilot-assisted transmission, and mismatch nearest-neighbor decoding. For example, a key enabler of URLLC is frequency diversity, which can be captured in our framework using a block-fading model in which each transmission frame spans a number of resource blocks in the time-frequency plane that are separated in frequency by more than the channel coherence bandwidth. The resulting channel is not memoryless, but rather block memoryless. Our analysis can be generalized to this setup by replacing the finite-blocklength bounds in Theorem 2 and 4 with their block-memoryless counterparts (see e.g., [35], [36]). Pilot-assisted transmission for channel estimation and mismatch nearest-neighbor decoding at the receiver can be also accounted for in the analysis [35], [36]. Note that the inclusion of mismatch nearest-neighbor decoding allows one to characterize the impact on performance of an

unknown interference signal, when such a signal is treated by the decoder as an additive noise.

B. Beyond the Point-to-Point Setup

Our analysis pertains a point-to-point link serving a single flow of data. The following are natural generalizations: (i) a point-to-point link serving multiple independent flows of information, possibly with different reliability, freshness, and latency requirements or priorities as in [37], [38]; (ii) a broadcast scenario with common information to be sent to all receivers; (iii) a broadcast scenario where independent information flows are to be sent to different receivers; (iv) the multiple-access case, where multiple sensors send information to a common destination; (v) the multi-hop setting where the objective is to maintain low end-to-end latency and high end-to-end data freshness.

The first three scenarios can be addressed by generalizing the framework presented in this paper as discussed below. For scenario (i), the steady-state average age of information under LCFS-S and LCFS-Q was derived in [39]. An analysis of the delay and the peak age violation probabilities could be tackled using the tools we presented in this paper. To analyze scenario (ii), one can leverage the nonasymptotic bounds presented in [40] on the performance of VLSF codes on the broadcast channel with common message. A special case of scenario (iii), in which at each time the transmitter chooses a single flow to serve, was recently analyzed in [41]. For the case of deterministic packet arrivals, one can leverage the indexability of the problem. For more general arrival distributions, the multi-server results in [42] on the optimality of the last-generated first-serve packet management policy can be leveraged. All these analyses rely on a simplified physical-layer model and can be combined with the framework introduced in this paper to obtain more accurate performance predictions, and derive guidelines for a joint design of physical and higher layers.

VIII. Conclusion

The optimal design of URLLC must rely on nonasymptotic joint coding-queuing analyses that target tail probabilities rather than averages. In this paper, we illustrated how to perform such an analysis for the case of i.i.d. Bernoulli packet arrivals, single-server queue, and point-to-point communications over a bi-AWGN channel via VLSF codes. We presented methods for evaluating the delay violation and the peak-age violation probability and analyze the dependence of these two quantities on system parameters such as the frame length, the undetected-error probability and the packet-management policy. A novel aspect of our approach is that it accounts also for undetected errors. Such errors may be extremely harmful in mission-critical communication systems and specific designs aimed at minimizing them will be the subject of future works.

Our analysis illustrates that the delay violation probability is extremely sensitive to the value of the system parameters. This can be seen in Figs. 5 and 7, which show the dependence of the maximum throughput and of the delay violation probability on the frame size for the case of simple ARQ with perfect error detection, and in Fig. 6, which depicts the dependence of the delay violation probability on the undetected error probability.

We also showed that analyses based on average latency or on a stochastic-network-calculus approach do not yield an optimal design and, for the case of stochastic network calculus, suffer from lack of generality.

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., $R_m = Q_{T_m+D_m}$. We observe that $\{R_m\}_{m=1}^{\infty}$, is a Markov chain governed by $R_{m+1} = \max\{R_m - 1, 0\} + U_m$ where $U_m = \sum_{t=1}^{S_{m+1}} \mathbf{1}\{B_{\max(T_m+D_m, T_{m+1})+t} > 0\}$ is the number of bulks of packets arriving during the service time of the mth bulk. Note that U_m depends only on the length of the service time S_{m+1} and is independent of R_m. Hence, $U_m \sim \sum_{t=1}^{S_1} \mathbf{1}\{B_t > 0\}$. The PGF of the steady-state buffer-size R is $G_R(s) = (1 - \mathbb{E}[U])(s - 1)G_U(s)/(s - G_U(s))$, $\mathbb{E}[U] < 1$ [30, Eq. (11.3.11)]. Since $R \sim \sum_{t=1}^{D} \mathbf{1}\{B_t > 0\}$, we have that $G_D(s) = G_R\left(G_{\mathbf{1}\{B_1>0\}}^{-1}(s)\right)$, where

$$G_{\mathbf{1}\{B_1>0\}}(s) = (1-\lambda)^n + s(1 - (1-\lambda)^n). \quad (38)$$

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

$$\mathbb{E}[U] = \mathbb{E}[\mathbf{1}\{B_1 > 0\}]\,\mathbb{E}[H_1]\,\mathbb{E}[N_1] = \lambda n\,\mathbb{E}[\tau]. \quad (39)$$

We observe that $G_U(s) = G_{S_1}\left(G_{\mathbf{1}\{B_1>0\}}(s)\right)$, $G_{H_1}(s) = G_\tau(s)$, and $G_{S_1}(s) = G_{N_1}(G_{H_1}(s))$ where

$$G_{N_1}(s) = ((1 - \lambda + \lambda s)^n - (1-\lambda)^n)/(1 - (1-\lambda)^n).$$

Algebraic manipulations yield (8).

APPENDIX B
COMPUTATION OF THE CCDF FROM THE PGF OF A RANDOM VARIABLE

We elaborate on the method used to compute the CCDF of a nonnegative integer-valued random variable X given its PGF $G_X(s)$. The chosen method is the recursion based z-transform inversion, which is exact, but it is applicable only when $G_X(s)$ is a rational function.

We start by noting that the z-transform of the CCDF can be written as

$$\sum_{k=0}^{\infty} s^k \Pr\{X > k\} = \frac{1 - G_X(s)}{1 - s}. \quad (40)$$

When $G_X(s)$ is a rational function, we can rewrite (40) as

$$\sum_{k=0}^{\infty} s^k \Pr\{X > k\} = \left(\sum_{k=0}^{n} a_k s^k\right) \Big/ \left(\sum_{k=0}^{m} b_k s^k\right) \quad (41)$$

for some positive integers m, n and real numbers $a_0, \ldots, a_n, b_0, \ldots, b_m$. Multiplying by $\sum_{k=0}^{m} b_k s^k$ both sides of (41) and equating the coefficients, we obtain

$$\Pr\{X > k\} = \frac{1}{b_0}\left(a_k - \sum_{u=1}^{k} b_u \Pr\{X > k - u\}\right)$$

where $a_k = 0$ for $k > n$ and $b_u = 0$ for $u > m$. The method is referred to as recursive because, in order to compute $\Pr\{X > k\}$, we need to compute first $\Pr\{X > u\}$ for all values of $u = 0, \ldots, k - 1$.

APPENDIX C
PROOF OF THEOREM 4

The choice of the auxiliary random variable U (see Def. 1), the random codebook construction, and the proof of (18) are the same as in the proof of [4, Thm. 3] and, hence, omitted. To prove (17), we first need to establish some notation. Let $C_1, \ldots, C_{2^k}$ be the 2^k random codewords. At time instant nt, $t = 1, 2, \ldots$ the decoder computes the 2^k information densities $\imath(C_w^{nt}; Y^{nt})$, $w \in \mathcal{W}$ and feeds back an ACK as soon as one of the information densities exceeds γ. Specifically, let $\tau_w = \inf\{t \geq 0 : \imath(C_w^{nt}; Y^{nt}) > \gamma\}$. The ACK is sent after frame $\tau = \min_{w \in \mathcal{W}} \tau_w$ and the output of the decoder is the message with index $\max\{w : \tau_w = \tau\}$. The CCDF of τ can be upper-bounded as

$$\Pr\{\tau \geq t\} = \frac{1}{2^k} \sum_{w \in \mathcal{W}} \Pr\{\tau \geq t \,|\, M = w\} \qquad (42)$$

$$\leq \frac{1}{2^k} \sum_{w \in \mathcal{W}} \Pr\{\tau_w \geq t \,|\, M = w\} \qquad (43)$$

$$= \Pr\{\tau^* \geq t\} \qquad (44)$$

where τ^* is defined in (15).

APPENDIX D
PROOF OF THEOREM 8

Let T be the number of frames elapsed between the arrival of two consecutive packets, which is geometrically distributed. Then, $G_T(s)$ takes the form given in (25). Note that under LCFS-S, preemption occurs with probability $1 - p_d$, where p_d is given in (36). Let F be the number of packets whose transmission is interrupted because of preemption, in between two successfully delivered packets. Then, given F, the peak age of information has the same distribution as the following sum of independent random variables: $A \sim T^{(0)} + \sum_{j=1}^{F} T_j^{(1)} + H^{(0)}$. Here, $T^{(0)}$ has the same distribution as T given that the first packet is delivered successfully, whereas $T_j^{(1)}$ has the same distribution as T given that the first packet is not delivered because of preemption. Furthermore, $H^{(0)}$ is the service time of a packet given that the packet is delivered (no preemption). One can show that $T_j^{(1)} \sim H^{(0)} \sim \mathrm{Geom}(1 - \epsilon_{\mathrm{rcus}}(1 - \lambda)^n)$, which implies (33), and also that $G_A(s) = \mathbb{E}\left[G_{T^{(0)}}(s)\left(G_{H^{(0)}}(s)\right)^{F+1}\right]$. To compute $G_{T^{(0)}}(s)$, we use that $G_T(s) = p_d G_{T^{(0)}}(s) + (1 - p_d)G_{T_j^{(1)}}(s)$ from which (37) follows. Finally, we obtain (35) by using that $F + 1 \sim \mathrm{Geom}(p_d)$.

REFERENCES

[1] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Delay and peak-age violation probability in short-packet transmission," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, U.S.A., Jun. 2018.

[2] ITU-R, "Recommendation ITU-R M.2083-0: IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," Tech. Rep., Sep. 2015.

[3] 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.

[4] ——, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[5] 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.

[6] İ. 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. D. Forney Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, Mar. 1968.

[14] 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.

[15] 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.

[16] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Delay performance of wireless communications with imperfect CSI and finite length coding," *IEEE Trans. Commun.*, 2018, to appear.

[17] 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.

[18] 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.

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

[20] 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.

[21] P. Larsson, J. Gross, H. Al-Zubaidy, L. K. Rasmussen, and M. Skoglund, "Effective capacity of retransmission schemes: A recurrence relation approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4817–4835, Aug. 2016.

[22] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[23] 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.

[24] 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.

[25] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 131–135.

[26] H. Sac, T. Bacinoglu, E. Uysal-Biyikoglu, and G. Durisi, "Age-optimal channel coding blocklength for a transmission queue with FCFS service and ARQ," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Kalamata, Greece, Jun. 2018.

[27] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random–coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, San Diego, CA, U.S.A., Feb. 2011.

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

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

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

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

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

[33] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 968–981, Sep. 1991.

[34] I. Rubin and M. Quaily, "Performance of finite capacity communication and queuing systems under various service and buffer preemption policies," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, New Orleans, LA, U.S.A., Mar. 1988, pp. 505–514.

[35] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-latency short-packet transmissions: Fixed length or HARQ?" in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Abu Dhabi, U.A.E., Dec. 2018.

[36] J. Östman, G. Durisi, E. G. Ström, M. C. Coşkun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Trans. Commun.*, 2018, to appear.

[37] E. Najm, R. Nasser, and E. Telatar, "Content based status updates," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, U.S.A., Jun. 2018, pp. 2266–2270.

[38] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sep. 2018.

[39] R. D. Yates and S. K. Kaul, "Status updates over unreliable multiaccess channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 331–335.

[40] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Common-message broadcast channels with feedback in the nonasymptotic regime: Stop feedback," *IEEE Trans. Inf. Theory*, vol. 64, no. 12, pp. 7686–7718, Dec. 2018.

[41] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, pp. 1–14, Oct. 2018.

[42] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Age-optimal information updates in multihop networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Aug. 2017, pp. 576–580.