Throughput Analysis of Low-Latency IoT Systems with QoS Constraints and Finite Blocklength Codes

Yulin Hu, *Senior Member, IEEE,* Yi Li, M. Cenk Gursoy, *Senior Member, IEEE,* Senem Velipasalar, *Senior Member, IEEE,* and Anke Schmeink, *Senior Member, IEEE*

Abstract

Internet of Things (IoT) is a promising paradigm to connect massive number of devices in future wireless communications while satisfying various quality of service (QoS) requirements. In this paper, we consider a QoS-constrained IoT system operating with finite blocklength (FBL) codes to support low latency communications. Two data arrival models are considered, namely, constant-rate arrival and ON-OFF discrete-time Markov arrival. The throughput performance is studied for both arrival models under statistical queuing constraints and deadline limits. For the scenario with instantaneous channel state information (CSI), we derive the QoS-constrained throughput expressions for both arrival models. Subsequently, an instantaneous-CSI-driven optimal power control algorithm is proposed to maximize the throughput, while guaranteeing a certain reliability target. In addition, we consider a scenario with only average CSI being available at the transmitter and propose to apply hybrid automatic repeat request (HARQ) schemes to improve the FBL performance. The decoding error probability and the outage probability are first characterized, following which the distribution of transmission period is derived. Furthermore, the throughput expressions are provided for both types of arrivals. Via numerical analysis, the impact of error probability, fixed transmission rate, coding blocklength, and QoS constraints on the throughput is studied.

Index Terms

Finite blocklength, HARQ, Internet of Things (IoT), Markov arrivals, power control, QoS.

I. Introduction

FUTURE wireless networks are expected to support high speed, low-latency and high reliability transmissions while connecting a massive number of smart devices, e.g., enabling the Internet of Things (IoT) [2], [3]. In particular, designers and researchers of IoT systems are increasingly interested in having wireless links supporting delay-sensitive data traffic generated in applications such as industrial control applications, autonomous driving, cyber-physical systems, E-health, and haptic feedback in virtual and augmented reality. The common features of these IoT applications [5]–[8] is that the transmission reliability is usually a concern, and more importantly that due to low latency constraints, the coding blocklengths for wireless transmissions are quite short.

Moreover, various IoT applications, such as streaming multimedia [9], augmented reality (AR) and online gaming, have certain quality of service (QoS) requirements in order to avoid excessive buffer overflows, data packet drops, and violations of delay constraint. For such delay sensitive applications, in [10] the effective capacity formulation was developed to characterize the capacity/throughput under statistical queuing constraints in the form of limitations on the asymptotic behavior of the buffer overflow probability. In particular, a large-deviations setting is considered, in which queueing constraints require the buffer overflow probability to decay exponentially fast for sufficiently large buffer thresholds.

Facilitated by the effective capacity model, many wireless communication schemes have been proposed recently to improve the QoS performance under certain latency constraints. In general, these techniques

The material in this paper has been presented in part at the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, in June 2017 [1]. This work was supported in part by the DFG research grant SCHM 2643/16.

Y. Hu and A. Schmeink are with ISEK Research Group, RWTH Aachen University, D-52074 Aachen, Germany (email: {hu, schmeink}@isek.rwth-aachen.de).

Y. Li is a wireless system engineer at Intelligent Fusion Technology, Germantown, MD 20876, US (email: yli33@syr.edu)

M. C. Gursoy and S. Velipasalar are with Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, US (email: {mcgursoy, svelipas}@syr.edu).

can be roughly considered as being motivated by the following two strategies. On the one hand, when the instantaneous channel state information (CSI) is available at the transmitter, optimal scheduling and resource allocation schemes are proposed to improve the QoS-constrained performance (e.g., effective capacity) according to the CSI. For instance, QoS-driven optimal power control policies are proposed for broadcast channels [11] and for multiple-access channels [12]. Under certain QoS constraints, the authors in [13] study the energy-efficient power control for 5G cellular networks. In addition, a joint power allocation and framework design is presented in [14] to maximize the effective capacity. In order to satisfy the QoS requirements, the optimal power allocation schemes are proposed for a multi-user network in [16] and a relay network [17]. More recently, the QoS-constrained resource allocation designs are introduced for a downlink multiple user network [18] and a multi-user network with device-to-device communications [19]. Moreover, for a QoS-constrained non-orthogonal multiple access network, a sub-optimal power control policy is proposed in [20].

On the other hand, when instantaneous CSI is not available at the transmitter, relatively more conservative approaches, e.g., reducing packet size to improve the reliability or shortening blocklength to allow more retransmissions, are generally preferred as they are expected to be more reliable. One promising approach is to apply automatic repeat request (ARQ) scheme, which is well-known as one of the key performance enhancement techniques for wireless transmissions systems without instantaneous CSI at the transmitter. For such systems, by retransmitting the erroneously decoded packets, ARQ schemes are able to improve the reliability as well as to adapt the average transmission data rate to the time-varying channel qualities. By combining the forward error correction mechanisms with ARQ, hybrid ARQ (HARQ) protocols are proposed, which can achieve higher reliability performance [21]. In particular, when the erroneously decoded packets are combined with incremental redundancy (IR) after each retransmission in HARQ, these HARQ schemes are called HARQ-IR, under which the decoding error probability is decreasing in the number of retransmissions [22], [23]. It should be pointed out that it is challenging to characterize the QoS-constrained performance of HARQ schemes due to their inherent retransmission process. With this motivation, the effective capacity of HARQ-IR have been derived for a system with reliability constraints [24] and for a system with a recurrence-relation retransmission approach [25].

However, in previous works addressing either power control or HARQ under queuing constraints, it was generally assumed that the instantaneous transmission data rates were given by the Shannon capacity, which is only true when the code blocklengths grow to be infinite long, i.e., so called the infinite blocklength (IBL) regime. Unfortunately, in a practical system the codes can only have finite blocklengths (FBL). In particular, for a QoS-constrained IoT network, the coding blocklengths are required to be quite short to satisfy the delay deadlines and queuing constraints. Hence, it is more essential to study the FBL performance while explicitly taking into account decoding error probabilities in the analysis of the QoS-constrained performance. This motivated us to leverage recent advances in the characterization of coding rates in the FBL regime [26]–[28], and study the FBL throughput performance of both instantaneous-CSI-driven power control and non-CSI-based HARQ-IR schemes, while all transmissions are subject to statistical queuing constraints at the transmitter buffer[1]. Moreover, in related works which investigating the effective capacity of either power control or HARQ-IR, it is generally assume that the data arrives at the transmitter with a constant rate. However, in addition to the constant arrivals, the randomly time-varying arrivals also need to be addressed (at least) for certain traffic types in IoT networks [32]. For instance, the data traffic can be modeled as an ON-OFF process in non-continuous sensor feedback/report, and the variable bit-rate traffic (e.g., video streams) is statistically characterized as autoregressive, Markovian, or Markov-modulated processes [33]. Therefore, in this work we also extend our throughput analysis on both instantaneous-CSI-driven power control and non-CSI-based HARQ-IR to a scenario with data arrivals

[1] It should be mentioned that most existing studies on FBL networks are aiming at characterizing/improving the physical-layer performances [29], [30], i.e., throughput and reliability. Nevertheless, a joint packet dropping and resource allocation policy has recently been proposed in the FBL regime to minimize the transmit power under delay and reliability constraints [31]. However, to the best of our knowledge, it is still an open problem to address the QoS-constrained throughput levels in the presence of constant and dynamic data arrivals (at the transmitter) in the FBL regime under instantaneous-CSI-driven and non-CSI-based transmission policies.

being modeled as ON-OFF discrete-time Markov processes.

The rest of the paper is organized as follows. In Section II, we first describe our system model. Subsequently, we introduce the QoS-constrained throughput which is the performance metric considered in this work, and then provide the problem statements addressed in the succeeding sections. We present our main results in Sections III through V. In particular, we analyze the instantaneous-CSI-driven optimal power control in Section III and also study the corresponding QoS-constrained throughput while assuming a constant data arrival source. In Section IV, we investigate the performance of the HARQ-IR in the absence of instantaneous CSI at the transmitter and again identify the QoS-constrained throughput. Section V extends the results in the preceding sections to a system with a random data arrival source. Finally, numerical results are provided in Section VI and we conclude our analysis in Section VII.

II. Preliminaries

In this section, we describe the interested reliable and low-latency IoT system and introduce preliminaries on statistical queuing constraints, and QoS-constrained throughput metrics.

A. System Description

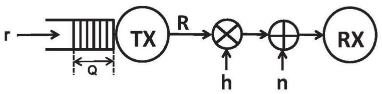

Fig. 1. An example of the considred system.

As shown in Figure 1, we consider a simple wireless IoT network with a buffered transmitter and a receiver. Data packets arrived at the transmitter are initially stored in a buffer waiting for being transmitted to the receiver, satisfying certain queuing constraints. The channel is assumed to experience block flat-fading. In other words, the channel fading stays constant within a block of l symbols, i.e., the code bocklength is assumed to be shorter than the channel coherence time, which matches with the low-latency requirements of IoT applications. In the i^{th} time block, the received signal at the receiver is given by:

$$\mathbf{y}_i = h_i \mathbf{x}_i + \mathbf{n}_i\,, \tag{1}$$

where $i = 1, 2, \ldots$. In addition, $\mathbf{x}_i$ and $\mathbf{y}_i$ are signal vectors of length l, representing the transmitted and received signals, respectively. h_i is the channel fading coefficient in the i^{th} block. $\mathbf{n}_i$ is the noise vector, which has independent and identically distributed zero-mean Gaussian components, each with variance N_0. The channel gain in the i^{th} time block is denoted by $z_i = |h_i|^2$. In addition, the channel gain in each block is assumed to have the probability density function $f_{\text{PDF}}(z)$. Moreover, denote by p_{tx} the per symbol transmit power, i.e., $p_{\text{tx}} = \mathbb{E}\{\|\mathbf{x}_i\|^2\}/l$. Therefore, the signal-to-noise ratio (SNR) at the transmitter can be obtained by $\text{SNR} = \frac{p_{\text{tx}}}{N_0}$.

B. Statistical Queuing Constraints and QoS-Constrained Throughput

Recall that in this work the transmitter is assumed to operate under a queuing constraint. According to [10], this requires the buffer overflow probability to decay exponentially fast, i.e.,

$$\Pr\{Q \geq q\} \approx \varsigma e^{-\theta q}, \tag{2}$$

for a sufficiently large overflow threshold q, where Q is the stationary queue length and $\varsigma = \Pr\{Q > 0\}$ is the probability that the buffer is not empty. In addition, θ is so-called the QoS exponent. More precisely, θ is defined in [34] by [2]

$$\theta = \lim_{q \to \infty} \frac{-\log \Pr\{Q \geq q\}}{q}. \tag{3}$$

[2] In the paper, we use logarithm expressed without a base, i.e., $\log(\cdot)$, to represent the natural logarithm $\log_e(\cdot)$.

Note that θ actually controls the exponential decay rate of the buffer overflow probability. In particular, according to (2), it can be noticed that a higher value of θ indicates a stricter limitation on the buffer overflow probability, which results in a more stringent QoS constraint. On the contrary, a small θ represents a relatively looser QoS constraint.

Note that the system operates in a time-slotted fashion, where time is divided into blocks with lengths of l symbols. A certain amount of data is transmitted in each block, during which we also potentially have new data arrivals into the transmitter buffer. We denote the instantaneous arrival rates (bits/block) in the arrival process by a_i and the departure rates (bits/block) in the departure process by c_i, respectively. According to the effective bandwidth [34] and effective capacity [10] formulations, the data arrival and departure processes at the buffer should satisfy the steady-state condition, i.e., the following equality holds

$$\Lambda_a(\theta) + \Lambda_c(-\theta) = 0, \tag{4}$$

where $\Lambda_s(\theta) = \lim_{t\to\infty} \frac{1}{t} \log_e \ \mathbb{E}\{e^{\theta \sum_{i=1}^{t} s_i}\}$ is the so-called asymptotic logarithmic moment generating function (LMGF) of the random process s_i. Equation (4) allows us to determine the average arrival rates for different departure and arrival models by formulating $\Lambda_a(\theta)$ and $\Lambda_c(-\theta)$ and substituting them into (4).

Under a constant arrival model, we have $a_i = la$ (bits/block) for all i, where a is the constant rate in bits/symbol and l is the blocklength. Clearly, it holds

$$\Lambda_a(\theta) = la\theta. \tag{5}$$

Hence, according to (4), we have

$$la = -\frac{1}{\theta}\Lambda_c(-\theta). \tag{6}$$

In particular, the right side of (6) is so-called the effective capacity (in bits/block) [10], denoted by C_E. The effective capacity characterize the maximum constant arrival rate that can be supported by the time-varying wireless transmission rates while satisfying the statistical queuing constraint in (2). When the arrival rate is random, the computation of the system throughput is more complicated. Generally, it is required to formulate the LMGF of the arrival process as a function of the average arrival rate, and then obtain the throughput via solving Equation (4).

C. Problem statements

In this work, we study the throughput of low-latency IoT communication systems operating with delay QoS constraints and finite blocklength codes. Hence, we consider both queueing delays (via the QoS exponent θ) and the transmission delay (via the use of finite blocklength codes). In particular, we analyze the low-latency throughput in two scenarios with different CSI assumptions. First, we consider a scenario in which the instantaneous CSI is assumed to be available at the transmitter. In this setting, we apply an optimal power control policy at the transmitter to maximize the QoS-constrained throughput under reliability constraints.

Secondly, we consider a scenario in which no transmitter CSI is assumed to be available and the transmitter sends the data at fixed rates. To guarantee the QoS of transmissions under this no instantaneous CSI scenario, we employ the HARQ-IR scheme and study the corresponding low-latency throughput achieved under delay QoS constraints with finite blocklength codes.

Moreover, extending our initial analysis assuming a constant data arrival source, we also address the QoS performance, considering a random data arrival source.

III. Instantaneous-CSI-Driven Power Control with Constant Data Arrival

In an additive white Gaussian noise (AWGN) channel, the normal approximation of the coding rate r (in bits per channel use) of transmission is derived in [26], [27]. Later on, this approximation is improved to become tighter in [28], where the third-order term of the normal approximation is derived. In particular, with error probability ν, transmitted signal-to-noise ratio SNR, channel gain z and blocklength l, the coding rate is shown to have the following asymptotic expression [28]:

$$\begin{aligned} r(z\text{SNR}, \nu) = & \log_2(1+z\text{SNR}) \\ & -\sqrt{\frac{z\text{SNR}(z\text{SNR}+2)}{l(z\text{SNR}+1)^2}} Q^{-1}(\nu)\log_2 e + \frac{\log l}{l} + \frac{o(1)}{l}, \end{aligned} \tag{7}$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ is the Gaussian Q-function.

Note that we consider a scenario with a fading channel, i.e., the channel quality varies over time. As the CSI is available, optimal power control schemes can be applied to maximize the QoS-constrained FBL throughput, given by [41]

$$C_{\text{E}} = -\frac{1}{\theta l} \log\left\{\mathbb{E}_z\left[e^{-\theta l r(z, \frac{p_{\text{tx}}}{N_0})}(1-\nu) + \nu\right]\right\}. \tag{8}$$

Obviously, for a given ν, C_{E} is influenced by $\{p_{\text{tx}}\}$. Our aim is to optimally allocate power over time/channel qualities to maximize the above throughput, while guaranteeing a target error probability ν, where ν is within a range of practical interest. Moreover, we consider the practical assumption that IoT applications generally require guarantees for a basic reliable connection/transmission as long as the channel state is not extremely bad, i.e., $z \geq z_{\text{min}}$. This requirement in terms of SNR is given by $\gamma = \frac{p_{\text{tx}} z}{N_0} \geq \gamma_{\text{th}} \geq 0$ dB, while the equivalent requirement in terms of coding rate is $r \geq r(\gamma_{\text{th}}, \nu)$. On the other hand, as the channel has a random behavior, it is possible that $z_i \leq z_{\text{min}}$. In this case, we simply skip this transmission, i.e., allocate zero power to the transmission in this fading block.

Therefore, the objective of the power control is to maximize the throughput by optimally allocating power over frames (time), while satisfying the average (over time) power constraint, i.e., $\mathbb{E}_i\{p_{\text{tx}}\} = \mathbb{E}_z\{p_{\text{tx}}\} \leq p_{\text{ave}}$. Hence, the optimal power level p_i^* in the i^{th} block depends both on the realization of the instantaneous channel gain z_i and its distribution, i.e., $f_{\text{PDF}}(z)$. Based on the above analysis, the optimization problem of power control (i.e., power allocation across time) in the presence of constant arrivals is stated as follows:

$$\begin{aligned} \max_{p_{\text{tx}} \in \Psi} \quad & C_{\text{E}} \\ s.t. \quad & \mathbb{E}_z\{p_{\text{tx}}\} \leq p_{\text{ave}}, \end{aligned} \tag{9}$$

where $\Psi = \begin{cases} p_{\text{tx}} \geq \frac{N_0 \gamma_{\text{th}}}{z}, \text{if } z \geq z_{\text{min}}, \\ p_{\text{tx}} = 0, \quad \text{if } z < z_{\text{min}}, \end{cases}$ is the feasible set of p_{tx}. Then, we have the following result for Problem (9).

Theorem 1. *Problem* (9) *is strictly convex when the blocklength and the target error probability are within practical interest, i.e.,* $l \geq 100, \nu \geq 10^{-24}$.

Proof: As the constraint is a linear function of the transmission power, Theorem 1 holds if C_{E} is concave in p_{tx} in the feasible set Ψ. In the following, we prove this concavity by showing $\frac{\partial^2 C_{\text{E}}}{\partial p_{\text{tx}}^2} \leq 0$, $\forall p_{\text{tx}} \in \Psi$. According to (8), we have

$$\frac{\partial C_{\text{E}}}{\partial p_{\text{tx}}} = \frac{\partial C_{\text{E}}}{\partial r} \frac{\partial r}{\partial \text{SNR}} \frac{\partial \text{SNR}}{\partial p_{\text{tx}}} = \frac{\partial C_{\text{E}}}{\partial r} \frac{\partial r}{\partial \text{SNR}} \cdot \frac{1}{N_0}, \tag{10}$$

$$\frac{\partial^2 C_{\text{E}}}{\partial p_{\text{tx}}^2} = \frac{1}{N_0^2} \frac{\partial^2 C_{\text{E}}}{\partial^2 r} \left(\frac{\partial r}{\partial \text{SNR}}\right)^2 + \frac{1}{N_0^2} \frac{\partial C_{\text{E}}}{\partial r} \frac{\partial^2 r}{\partial \text{SNR}^2}. \tag{11}$$

According to (8), we also have

$$\frac{\partial C_{\mathrm{E}}}{\partial r} = \frac{l e^{-\theta l r}(1-\nu)}{\mathbb{E}_z[e^{-\theta l r}(1-\nu)+\nu]} \geq 0, \tag{12}$$

$$\frac{\partial^2 C_{\mathrm{E}}}{\partial r^2} = \frac{-\theta l^2 e^{-\theta l r}(1-\nu)\nu}{\mathbb{E}_z[e^{-\theta l r}(1-\nu)+\nu]^2} \leq 0. \tag{13}$$

Hence, for $p_{\mathrm{tx}} \in \Psi$, $\frac{\partial^2 C_{\mathrm{E}}}{\partial p_{\mathrm{tx}}^2} \leq 0$ holds if $\frac{\partial^2 C_{\mathrm{E}}}{\partial \mathrm{SNR}^2} \leq 0$.

$$\frac{\partial^2 r}{\partial \mathrm{SNR}^2} = \frac{z^2 \log e}{(1+z\mathrm{SNR})^3}\left\{\phi - (1+z\mathrm{SNR})\right\}, \tag{14}$$

where $\phi\left(z\mathrm{SNR}\right) = \frac{A}{2\left(z^2\mathrm{SNR}^2+2z\mathrm{SNR}\right)^{\frac{3}{2}}} + \frac{3A}{\sqrt{z^2\mathrm{SNR}^2+2z\mathrm{SNR}}}$ and $A = Q^{-1}(\nu)\sqrt{\frac{1}{l}}$. When $\nu \geq 0.5$, we have $A \leq 0$. Then, $\frac{\partial^2 r}{\partial \mathrm{SNR}^2} < 0$. On the other hand, when $\nu < 0.5$, $\frac{\partial^2 r}{\partial \mathrm{SNR}^2}$ is increasing in A and therefore decreasing in ν and l. For an extreme scenario where $l = 100, \nu = 10^{-24}$, we have $A = 1.0199$. Then, $\frac{\partial^2 r}{\partial \mathrm{SNR}^2} \leq 0$ if $\phi\left(z\mathrm{SNR}\right) \leq 0$. Obviously, $\phi\left(z\mathrm{SNR}\right)$ is decreasing in $z\mathrm{SNR}$ for $A > 0$. In particular, we have $\phi\left(1\right) = -0.0372$ for $A = 1.0199$. Hence, $\phi\left(z\mathrm{SNR}\right) < 0$ for $z\mathrm{SNR} \geq 1 = 0$ dB. Note that the feasible set of p_{tx} definitely satisfies $z\mathrm{SNR} \geq 1 = 0$ dB, thus Problem (9) is a strictly convex optimization problem[3]. ■

According to Theorem 1, A similar problem in the IBL regime has been addressed in [35], Problem (9) can be solved via the Lagrange dual method[4]. Denote by λ the Lagrange multiplier associated with the average power constraint. The partial Lagrange dual function of Problem (9) is given by

$$L = -C_{\mathrm{E}} + \lambda\,\mathbb{E}_z\left\{p_{\mathrm{tx}} - p_{\mathrm{ave}}\right\}. \tag{15}$$

To obtain the dual optimal, we let $\frac{\partial L}{\partial p_{\mathrm{tx}}(z)} = 0$. Let us introduce a function

$$g = e^{-\theta C_{\mathrm{E}}} = \mathbb{E}_z\left\{e^{-\theta l r}(1-\nu)+\nu\right\}, \tag{16}$$

and express $C_{\mathrm{E}} = -\frac{1}{\theta}\log g$. Then, $\frac{\partial L}{\partial p_{\mathrm{tx}}(z)}$ is given by

$$\frac{\partial L}{\partial p_{\mathrm{tx}}(z)} = \varphi\left(1 - \frac{A}{\sqrt{a}}\left(1-a\right)\right)(1-a)^{\frac{\eta l+1}{2}} e^{\eta l A \sqrt{a}} - \lambda = 0, \tag{17}$$

where $\varphi = \frac{z(1-\nu)\eta l^{-\eta}}{g\theta N_0} f_{\mathrm{PDF}}\left(z\right)$, $\eta = \frac{\theta}{\log 2}$ and $a = 1 - \left(1 + \frac{p_{\mathrm{tx}} z}{N_0}\right)^{-2}$.

Since the dual problem is always convex, $\frac{\partial L}{\partial p_{\mathrm{tx}}(z)}$ is monotonically increasing in p_{tx}, and it is easy to solve (17) within the feasible set Ψ, and determine λ and p_{tx}^*. However, it should be pointed out that as p_{tx}^* and λ are generally interdependent on each other, it is unlikely to obtain a closed-form expression for the optimal power control policy. But the optimal power control can be obtained numerically. In the following, we propose an iterative algorithm for determining the optimal instantaneous transmit power based on both the instantaneous channel gain and its distribution. The key idea of the algorithm is to first initialize the values of λ and g, and obtain the corresponding p_{tx} according to (17). Subsequently, we update g based on the obtained p_{tx} till g converges to a constant g°. Finally, we keep updating λ till (17)

[3] We note that general formulations and conditions to establish the convexity of Problem (9) are provided above in the proof of Theorem 1, and these can be used to establish the range of zSNR values for given coding blocklength m and target error probability ν. Indeed, we can easily show that for more practical scenarios, the concavity holds for even lower zSNR bounds: e.g., zSNR intervals in which convexity is satisfied are *i.* $[-3.1\text{ dB}, \infty)$ for $l = 100, \nu = 10^{-10}$, *ii.* $[-10.81dB, \infty)$ for $l = 200, \nu = 10^{-10}$, *iii.* $[-15.53dB, \infty)$ for $l = 300, \nu = 10^{-5}$.

[4] It should be pointed out that a similar problem has been addressed in [35] following the Shannon capacity, where the impact FBL is not considered. More recently, the convexity in Theorem 1 has been stated in [36], by considering an approximated effective capacity. Hence, a rigorous proof of the convexity of the effective capacity in the FBL regime with respect to the transmit power has not been pursued in this study.

Algorithm 1 for Optimal Power Control.

```
Determining the optimal power control policy
    a) Given $\lambda$, $g$, obtain $p_{\mathrm{tx}}$ according to (17).
    b) According to (16), update $g$ based on the PDF of $z$ and the obtained $p_{\mathrm{tx}}$.
    c) Based on the updated $g$, update $p_{\mathrm{tx}}$ according to (17).
    d) if the updated $p_{\mathrm{tx}}$ is out of the feasible set, jump to Step k).
    e) if $g$ converges to a constant $g^{\circ}$
    f)    then Along with the obtained $g^{\circ}$, we have $p_{\mathrm{tx}}^{\circ} = \max\{p_{\mathrm{tx}}, \frac{\gamma_{\mathrm{th}} N_0}{z}\}$, $\lambda^{*} = \lambda$.
    g)    else go back to Step c).
    h) Check if the obtained $p_{\mathrm{tx}}^{\circ}$ satisfies the average power constraint.
    i) if not satisfied with equality
    j)    then Update the value of $\lambda$ and go back to Step a);
    k)    else We have obtained the optimal power control policy, including $\lambda^{*}$ and the converged $g^{\circ}$.
Instantaneous power control at $i^{\mathrm{th}}$ block, $i = 1, 2, ...$
    l) Substituting the instantaneous $z_i$ as well as $\lambda^{*}$ and $g^{\circ}$ to (17), the optimal power $p_{\mathrm{tx}}^{*}$ is determined.
```

is satisfied. As the Problem (9) is strictly convex in the feasible set, the converged solution is optimal.

In particular, Algorithm 1 has two parts: In the first part, the optimal power allocation policy is fully characterized by the optimal value λ^{*} and the converged g°. Following these characterizations, the optimal power can be determined in the instantaneous power control (second) part, by simply plugging the instantaneous channel gain z, λ^{*} and g° into (17) and solving it, which has very low computational complexity. Note that the first part is only required to be computed once for a given channel distribution, while the second part has to be processed for each instantaneous channel. Overall, computational complexity of Algorithm 1 is low.

IV. Throughput with HARQ-IR in the Presence of Constant Data Arrivals

In this section, we assume that only the transmit SNR and the distribution of channel fading is known at the source. Under this scenario, an efficient way to guarantee the QoS of transmissions is to let the system employ HARQ-IR scheme. Recall that l is the blocklength of each codeword (equivalently the length of each fading block). Under the constant data arrival model, the transmission rate is lR (bits/block) at the transmitter, where R denotes the date rate in bits/symbol. We consider a deadline constraint l_{tot} in the HARQ to control the packet delay, i.e., the deadline constraint limits the maximum duration of a transmission period as M time blocks, considering the costs of each blocklength and the corresponding feedback. In particular, each data packet is encoded into M codeword blocks, while each blocklength is l symbols. In each time block, a data packet via a codeword block is transmitted from the transmitter to the receiver. If the receiver decodes it successfully, it reliably sends an acknowledgment (ACK) to the transmitter via an error-free feedback link, and a new data packet will be sent in the next time block. On the other hand, if the receiver fails to decode the data packet, a retransmission request is sent through the feedback link, and another codeword block carrying the same data packet will be sent in the next time block [21]. For simplicity, we assume that the ACK and the retransmission request have the same delay cost of l_o symbols. Hence, for a given deadline constraint l_{tot}, M is limited by $\lfloor l_{\mathrm{tot}}/(l + l_o) \rfloor$, where $\lfloor \cdot \rfloor$ is a floor function.

Note that the deadline constraint is considered. A transmission period is finished if the receiver decodes the packet correctly or if the deadline is reached. Then, the transmitter move on the next transmission period sending another data packet. When the deadline is reached while the packet is still has not been decoded, i.e., decoding error of this packet occurs M times, an outage happens. The outdated packet is dropped in such a case. We denote the random transmission period by T, i.e., $1 \leq T \leq M$, and denote by ε the outage probability.

$$\Pr\{T{=}t\}{=}\begin{cases} 1-\mathbb{E}_{\mathbf{z}}\left\{\nu_1\right\} & \text{for } t=1,\\ \Pr\{T\leq t\}-\Pr\{T\leq t-1\}{=}1-\mathbb{E}_{\mathbf{z}}\left\{\nu_t\right\}-\left(1-\mathbb{E}_{\mathbf{z}}\left\{\nu_{t-1}\right\}\right){=}\mathbb{E}_{\mathbf{z}}\left\{\nu_{t-1}\right\}-\mathbb{E}_{\mathbf{z}}\left\{\nu_t\right\} & \text{for } 2\leq t\\ \mathbb{E}_{\mathbf{z}}\left\{\nu_{M-1}\right\} & \text{for } t=M.\end{cases} \tag{20}$$

For the departure process, the rate is instantaneous. When a transmission period is finished at the i^{th} block, the departure rate is given by $c_i = lR$ (bits/block), otherwise, $c_i = 0$. It should be pointed out that $c_i = lR$ holds also when the packet is dropped, as it also contribute to the departure rate in the queuing analysis, i.e., the queue length is reduced no matter the packet is transmitted or dropped. The packet drop probability is actually the outage probability, which is incorporated after solving the average arrival rate from (4), and the throughput r_{th} is given by the maximum average arrival rate r_{avg} multiplied by $(1-\varepsilon)$, because only $(1-\varepsilon)$ fraction of the packets are received by the receiver, i.e., ε fraction of the packets are discarded on average.

In the following, we study the outage probability of HARQ-IR and subsequently address the throughput performance.

A. Outage Probability for HARQ-IR at Finite Blocklengths

When there is no instantaneous CSI to guide the system operation, an efficient way to guarantee the reliability of transmissions is to let the system employ the HARQ-IR scheme. Recall that under the HARQ-IR scheme, the receiver accumulates the received information in each transmission period. In particular, at the end of the m^{th} trial of the transmission period, in total m received codeword blocks has been combined at the receiver for decoding the packet. From the perspective of achievable rate, this is actually equivalent to decoding one codeword with m subblocks while each length of subblock is l symbols. According to the FBL coding rate model in [26], the fixed transmission rate is given by

$$\begin{aligned} R = & \sum_{i=1}^{m}\log_2\left(1+z_i\text{SNR}\right)+\frac{\log(l)}{l}+\frac{o(1)}{l}\\ & -\sqrt{\sum_{i=1}^{m}\frac{(z_i\text{SNR}+2)z_i\text{SNR}}{l(z_i\text{SNR}+1)^2}}Q^{-1}(\nu)\log_2 e \end{aligned} \tag{18}$$

for the m^{th} trial, where ν is the decoding error probability, l is the blocklength, and $z_i = |h_i|^2$ is channel gain. Based on (18), the decoding error probability of the m^{th} trial/attempt of the packet transmission can be expressed as

$$\nu_m(\mathbf{z}) = Q\left(\frac{\sum_{i=1}^{m}\log_2\left(1+z_i\text{SNR}\right)+\frac{\log(ml)}{l}-R}{\log_2 e\sqrt{\sum_{i=1}^{m}\frac{(2+z_i\text{SNR})z_i\text{SNR}}{l(z_i\text{SNR}+1)^2}}}\right) \tag{19}$$

for given channel gains $\mathbf{z} = (z_1, \cdots, z_m)$. Note that the duration of a transmission period is repreanted by T. Hence, the probability mass function (pmf) of T is given in (20) at the top of the next page.

Recall that it is assumed that in the m^{th} trial, if the receiver fail to decode the packet by using all these m codeword blocks, this confirms it cannot decode the packet by only the first $m-1$ trials. Thus, the probability that the receiver correctly decodes a packet within t trials can be given by

$$\Pr\{T\leq t\} = 1-\mathbb{E}_{\mathbf{z}}\left\{\nu_t\right\}. \tag{21}$$

According to (20), when $2 \leq t \leq M-1$, $T = t$ actually indicates that only the t^{th} trial has been decoded correctly and that the first $t-1$ trials have finished and were unsuccessful. In addition, when $T = M$,

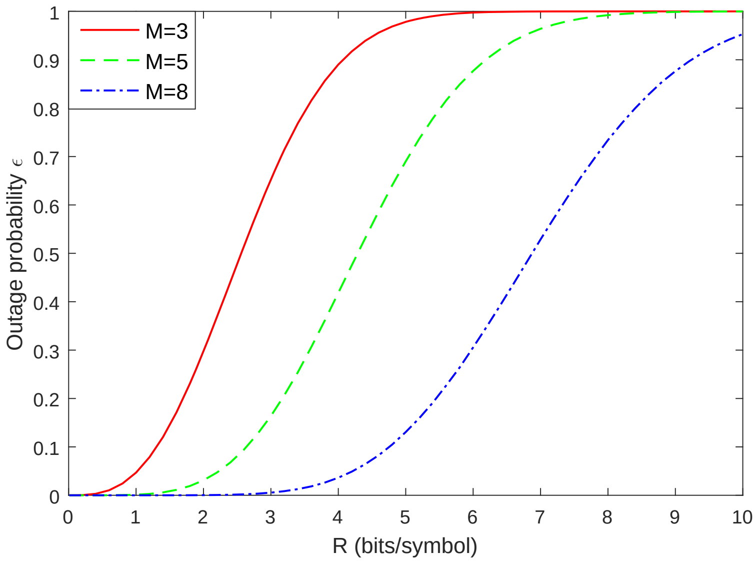

Fig. 2. Then relationship between R and the outage probability.

it indicates that the first $M-1$ trials have failed, while the last attempt does influence the pmf due to the fact that the deadline constraint forces the transmission period stop anyway after M trials. Recall that when the decoding error of one packet occurs M times, outage happens. The outage probability is given by

$$\varepsilon = \mathbb{E}_{\mathbf{z}}\left\{\nu_M\right\}. \tag{22}$$

We plot in Fig. 2 the impact of fixed transmission data rate R on the outage probabilities under different deadline constraints M. In the figure, we set the blocklength to $l = 100$ and SNR to 0 dB. Clearly, as the decoding error probability ν is increasing in R, a large R leads to a higher outage probability. In addition, it is also observed that lower outage probabilities are obtained when M is relatively large (representing a loose deadline constraint). Moreover, the figure also demonstrates that the fixed transmission rate R can be numerically determined using (19) and (22) as long as a target decoding error probability ε and deadline constraint M are specified.

B. Throughput of HARQ-IR with Constant-rate Arrivals

Under the constant-rate arrival assumption, for single transmission (without HARQ), the throughput (in bits/symbol) is given by $(1-\varepsilon)$ times the effective capacity (normalized by the blocklength l):

$$r_{\text{th}} = (1-\varepsilon)C_E(\theta, \text{SNR})/l = -\frac{1-\nu}{l\theta}\Lambda_c(-\theta). \tag{23}$$

Then, we have the following proposition addressing the throughput of HARQ-IR

Proposition 1. *Under constant-rate arrivals, with fixed transmission rate R (bits/symbol), given QoS exponent θ and specified deadline constraint M, the throughput of HARQ-IR scheme is given by*

$$r_{th} = (1-\varepsilon)C_E/l\,, \tag{24}$$

where C_E is the effective capacity (in bits/block), given by

$$C_E = -\frac{1}{\theta}\log\left(\max\left\{|\beta_1|, \cdots, |\beta_M|\right\}\right), \tag{25}$$

where $\{\beta_1, \cdots, \beta_M\}$ are the eigenvalues of the matrix $\mathbf{A}$ given in (26) at the top of the next page.

By applying Theorem 1 of reference [25] to the model considered in our work, Proposition 1 is verified. Note that the packet drop is not considered in the model of [25]. Hence, to apply the theorem in our model, the variable ν in [25] needs to be redefined as the number of packets leaving the queue in a

$$\mathbf{A} = \begin{pmatrix} \Pr\{T=1\}e^{-\theta lR} & \Pr\{T=2\}e^{-\theta lR} & \cdots & \Pr\{T=M-1\}e^{-\theta lR} & \Pr\{T=M\}e^{-\theta lR} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tag{26}$$

transmission period, which is always equal to 1 in our model due to the packet drop mechanism. In Section VI, simulation results are provided to verify our effective capacity characterization.

V. FBL Performance with ON-OFF Discrete-Time Markov Source

In this subsection, we analyze the throughput with finite blocklength codes when we have discrete-time Markov sources, with only two states, namely, ON and OFF states. In particular, we derive the maximum average arrival rate r_{avg} that can be supported by the wireless transmissions under statistical queuing constraints.

Under the discrete-time Markov model, the source keeps silent in the OFF state. On the other hand, data arrivals with rate $a_i = lr$ (bits/block) in the ON state, in which r (bits/symbol) is the constant rate in the ON state. Let us define states 1 and 2 as the OFF and ON states, respectively. The state transition probability matrix of this Markov source is given by

$$\mathbf{G} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \tag{27}$$

where p_{11} is the probability of the source staying in the OFF state in the next time block, while p_{22} is the probability of it staying in the ON state. In addition, p_{12} and p_{21} denote the probabilities that the source changes to a different state in the next time block. According to the properties of Markov processes, the probability of the ON state is given by

$$P_{ON} = \frac{1-p_{11}}{2-p_{11}-p_{22}}. \tag{28}$$

Therefore, we can derive the average arrival rate (in bits/symbol) of this ON-OFF Markov source by

$$r_{\text{avg}} = rP_{ON} = r\frac{1-p_{11}}{2-p_{11}-p_{22}}. \tag{29}$$

Note that the departure process and arrival processes are independent, the effective capacity characterizations in (8) and (25) are still valid for the random arrivals. Based on these two equations, the FBL throughput with a discrete-time Markov source is characterized in the following theorem

Theorem 2. *With the ON-OFF discrete-time Markov source, for given fixed transmission rate R (bits/symbol), QoS exponent θ and specified deadline constraint M, the throughput of HARQ-IR scheme is given by*

$$r_{th} = \frac{1-\varepsilon}{l}\frac{P_{ON}}{\theta}\log\left(\frac{e^{2\theta C_E} - p_{11}e^{\theta C_E}}{1-p_{11}-p_{22}+p_{22}e^{\theta C_E}}\right), \tag{30}$$

where C_E is the effective capacity given in (25).

Proof: According to [42] and (23), the LMGFs of the arrival process and the departure process at the transmitter can be by

$$\begin{cases} \Lambda_a(\theta) = \log_e\left(\frac{p_{11}+p_{22}e^{lr\theta}+\sqrt{(p_{11}+p_{22}e^{lr\theta})^2-4(p_{11}+p_{22}-1)e^{lr\theta}}}{2}\right) \\ \Lambda_c(-\theta) = -\theta C_E \end{cases}. \tag{31}$$

Applying the characterizations in (31) into (4), we have

$$lr = \frac{1}{\theta} \log \left(\frac{e^{2\theta C_E} - p_{11} e^{\theta C_E}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_E}} \right). \tag{32}$$

Plugging the above equality into (29), the maximum average arrival rate (in bits/symbol) is obtained as

$$r_{\text{avg}} = \frac{P_{ON}}{l\theta} \log \left(\frac{e^{2\theta C_E} - p_{11} e^{\theta C_E}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_E}} \right). \tag{33}$$

At the end of Section IV we have discussed that the throughput is given by $(1-\varepsilon) r_{\text{avg}}$. Therefore, Theorem 2 is verified. ■

Following the same steps as in the proof of Theorem 2 above, we can also prove that when instantaneous CSI is available at the transmitter and power control is employed, the throughput in the presence of an ON-OFF discrete-time Markov source is given by

$$r_{\text{th}} = \frac{P_{ON}}{l\theta} \log \left(\frac{e^{2\theta C_E} - p_{11} e^{\theta C_E}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_E}} \right), \tag{34}$$

where the effective capacity is now given by (8).

From (30) and (34), we can readily show that with ON-OFF discrete time Markov sources the throughput achieved with either the instantaneous-CSI-driven power control or the HARQ-IR is an increasing in the effective capacity C_E, which can be quickly verified by checking the first order derivative of r_{th} to C_E. Thus, the transmission parameters, including optimal power levels, fixed transmission rates r and outage probability ε, which are optimal solutions maximizing the throughput for the constant-rate arrival models, also optimize the throughput for the ON-OFF discrete time Markov arrival models.

The influence of the source burstiness was analyzed in [43], which showed that under queuing constraints the source burstiness degrades the energy efficiency. A similar discussion can be applied to the scenarios considered in our work. In particular, if the source remains in the ON state for a longer period, this indicates that it is less bursty. In this case, a smaller instantaneous arrival rates r is introduced for the fixed average arrival rate r_{avg}. As a results, if different sources have the same r_{avg}, then the one with less burstiness (corresponding to a smaller instantaneous arrival rate r) is more preferred in order to satisfy the queuing constraints.

VI. Simulation Results

The major scope of this work is to propose delay-aware system designs for improving the throughputs of low-latency communications. In particular, we consider two different scenarios. We identify the optimal power control under a scenario in which instantaneous CSI is available at the transmitter, while we apply the HARQ-IR scheme in the absence of CSI. In this section, we further investigate the corresponding throughput performances via numerical analysis and Monte Carlo simulations. In all the numerical analysis, we consider a Rayleigh channel fading model, where mean value of channel gain is $\mathbb{E}\{z\} = 1$. In addition, we set the average SNR at the receiver to 15 dB for the good average channel quality scenario and to 0 dB for the poor channel scenario.

A. *Throughput with instantaneous-CSI-driven optimal power control*

We start with the comparison between the proposed optimal power control scheme and constant transmit power schemes (without power control). Note that the proposed scheme guarantees the reliability ν instantaneously. In this comparison, we consider two schemes without power control: One scheme instantaneously guarantees the reliability by coding rate adaption while the other one just sets a constant coding rate as $r = 2.5$ bits/symbol. It should be mentioned that under the constant coding rate scheme, the coding rate is constant in time or channel fading, and therefore does not provide any guarantees regarding the reliability. The comparison results are shown in Fig. 3. First of all, we observe that the

QoS-constrained throughputs of schemes guaranteeing the reliability are quasi-concave in the (target) error probability ν, while the scheme with constant coding rate is not influenced by ν (as it does not provide reliability guarantees). In addition, the proposed optimal power control provides the highest throughput (by determining the optimal ν) in comparison to both schemes without power control. Moreover, comparing the two constant power schemes, we observe that the performance loss in terms of throughput is introduced by instantaneously guaranteeing the reliability.

In Fig. 4, we show the average throughput performance of the optimal power control with constant arrivals and ON-OFF Markov source, while varying the (target) error probability ν that needs to be satisfied for each instantaneous transmission. First of all, again confirming the results of Fig. 3, we observe that the QoS-constrained throughputs of both types of sources are quasi-concave in the (target) error probability ν. On the one hand, to satisfy an extremely low target error probability, the source has to determine a low coding rate, which leads to a low throughput. On the other hand, if the error probability is too high, it indicates that with a high probability the data packet needs to be retransmitted. These retransmissions cost additional time/symbol resources and increase queuing delays, thus reducing the QoS-constrained throughput. This characteristic of concavity indicates that the balance is achieved and the throughput is maximized at an optimal value of ν. In particular, when the target error probability ν is relatively low, it is preferred to simply apply this ν in (17) in the process of determining the optimal power control. However, if the required error probability of the service is relatively high, e.g., $\nu = 0.1$, it is preferred to set ν to be lower to maximize the QoS-constrained throughput. Secondly, as expected, a loose QoS requirement (corresponding to a low value of θ) results in a high throughput for both the scenarios with constant and random data arrival sources. Finally, in comparison to the constant data source, certain performance loss is introduced if the data arrival becomes ON-OFF Markov distributed. In particular, this performance loss is relatively small for the scenario with loose QoS requirements (low θ), while it is quite significant when the QoS requirement is strict (high θ).

Next, we investigate in more detail the impact of θ on the throughput performance of the proposed power control, for given (target) error probabilities. As shown in Fig. 5, all the throughput curves are strictly decreasing in θ. In addition, we observe a significant performance loss when the arrivals are random (compared to constant-rate arrivals) or when the blocklength is increased from 100 symbols to 1000 symbols. In other words, beside the data arrival model, the blocklength also influences the throughput performance. We further show in Fig. 6 the impact of the blocklength on the throughput. With a relatively large θ and small ν, the throughputs of (both constant and random rate arrivals) are decreasing in the

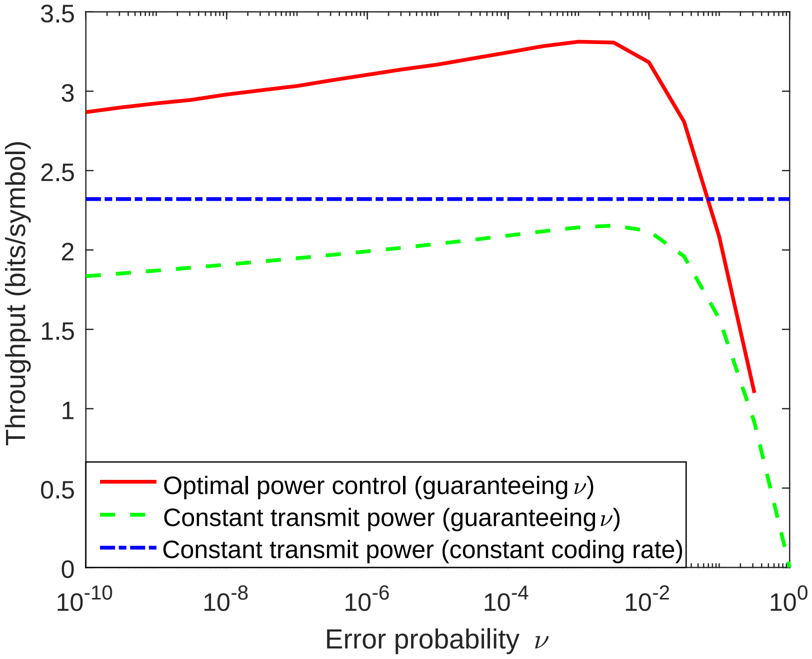

Fig. 3. The comparison between the proposed optimal power control and constant transmit power schemes.

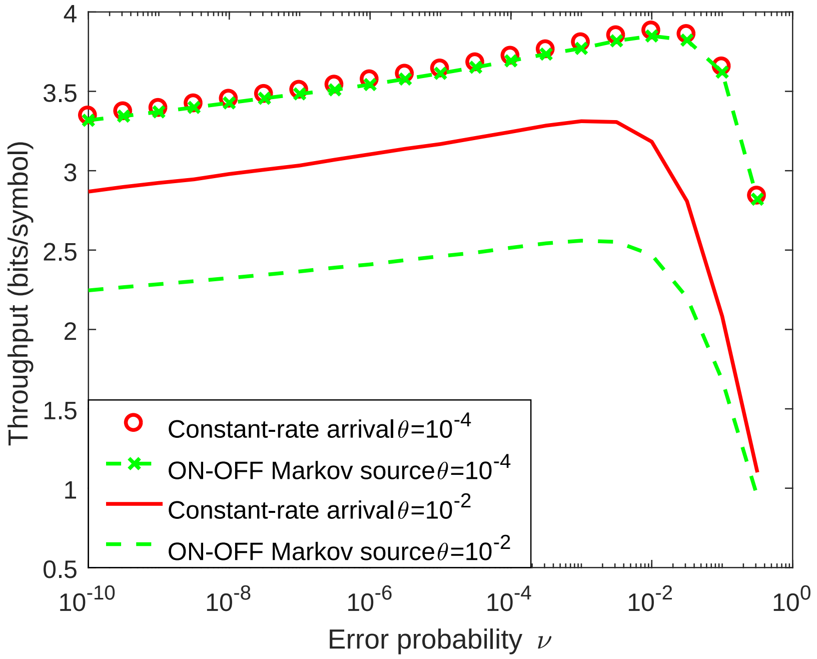

Fig. 4. The impact of decoding error probability ν on the throughput.

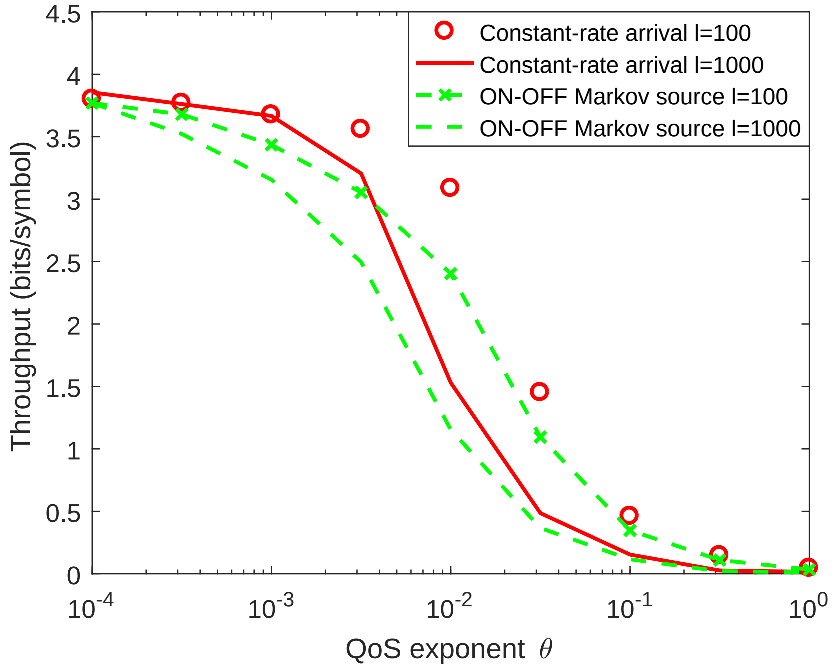

Fig. 5. The impact of QoS exponent θ on the throughput.

blocklength. On the other hand, it is observed that when θ is relatively small but ν is relatively large, the throughputs are concave in the blocklength. In particular, in this concave case, the throughputs are not very sensitive to the choice of blocklength. Hence, a short blocklength is in general a good choice for the QoS-supporting system with the proposed instantaneous-CSI-driven power control.

B. Throughput of HARQ-IR

In this subsection, we investigate the throughput performance of HARQ-IR in the scenario without instantaneous CSI feedback. We start with Fig. 7 to verify our analytical model in Proposition 1 and Theorem 2. In particular, we show in the figure the relationship between the buffer overflow threshold q and the logarithmic buffer overflow probabilities $\log \Pr\{Q \geq q\}$. In particular, the arrival rate for the constant-rate arrival model is given by the effective capacity in (25). For the ON-OFF discrete-time

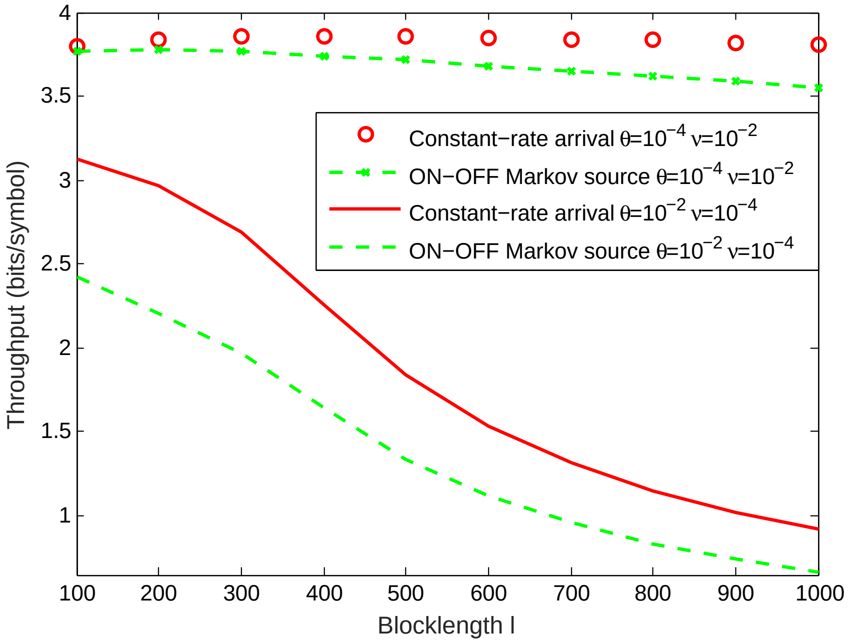

Fig. 6. The impact of blocklength l on the throughput.

Markov arrivals, we set $p_{11} = 0.3$, $p_{22} = 0.7$, and the arrival rates in the ON state are given by lr (bits/block) in (32). In Fig. 7, the curve of $\log(\Pr\{Q \geq q\}) = -\theta q$ is also provided as a reference. Note that $\Pr\{Q \geq q\} \approx \varsigma e^{-\theta q}$, where $\varsigma = \Pr\{Q > 0\}$ is the probability that the buffer is not empty. This reference curve actually indicates a lower bound representing the case of $\varsigma = 1$. As shown in the figure, since for either the constant-rate or ON-OFF Markov source the non-empty buffer probability ς is lower than 1, certain gaps exist between these two curves and the reference curve $\log(\Pr\{Q \geq q\}) = -\theta q$. More importantly, the logarithmic buffer overflow probabilities decrease almost linearly when q is sufficiently large, which matches well with the characterizations in (2) and (3)[5]. When $q > 1100$, by estimating the slopes of these two curves via linear regression, we find that the slopes are -0.0099 and -0.0100 for the constant-rate and ON-OFF Markov arrival models, respectively. Note that in the simulation we set $\theta = 0.01$, which corresponds to a slop of $-\theta = -0.01$. Hence, the slope errors of the two curves (in comparison to $-\theta$) are actually less than 1%, which is tiny, demonstrating the accurateness of our characterizations on the FBL throughput.

Next, we study the impact of the fixed transmission rate R on the throughput with constant arrival sources. The results are provided in Fig. 8 where we set the QoS exponent to $\theta = 0.1$. It is observed from Fig. 8 that the throughput is quasi-concave in R, i.e., there exists a globally optimal R maximizing the throughput. The explanation is as follows. On the one hand, for a small R, the departure rate is also low, i.e., this becomes the major limit (bottleneck) of the performance. On the other hand, when R is too large, it introduces a significant outage probability ε, and therefore a considerable fraction of the packets are dropped due to the deadline constraint. In this case, the decoding error probability becomes the bottleneck. We also learn from the figure that when M is relatively larger, which corresponds to a looser deadline constraint, a higher maximum throughput is achieved (by choosing the corresponding optimal value of R). Similar results were observed in [44] for small θ values in the IBL regime, i.e., without considering impact of finite blocklength. In particular, it has been shown in [44] that the throughput of HARQ-IR is increasing in the fixed transmission rate R as long as the deadline constraints are ignored. Therefore, our results are consistent with [44] in that a large M provides a high R and a low outage probability, thus resulting in a high throughput.

Finally, we investigate the influence of the finite blocklength l on the throughput for constant arrival models The results are shown in Fig. 9, where we set $\theta = 0.1$. It is observed that the throughput is

[5] Note that as long as (2) is true, $\log \Pr\{Q \geq q\} \approx -\theta q + \log \varsigma$ holds.

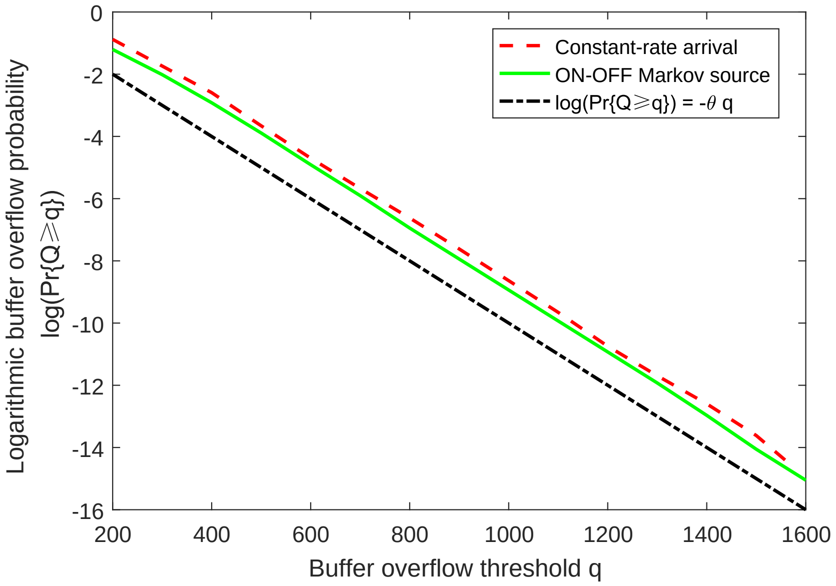

Fig. 7. The relationship between the buffer overflow threshold and the logarithmic overflow probability . In the numerical analysis, we set $\theta = 0.01$, $M = 5$, $R = 3$ (bits/symbol) and $l = 100$. For each curve, we repeat the simulations 2000 times, and in each time the simulation is conducted over 10^5 time blocks.

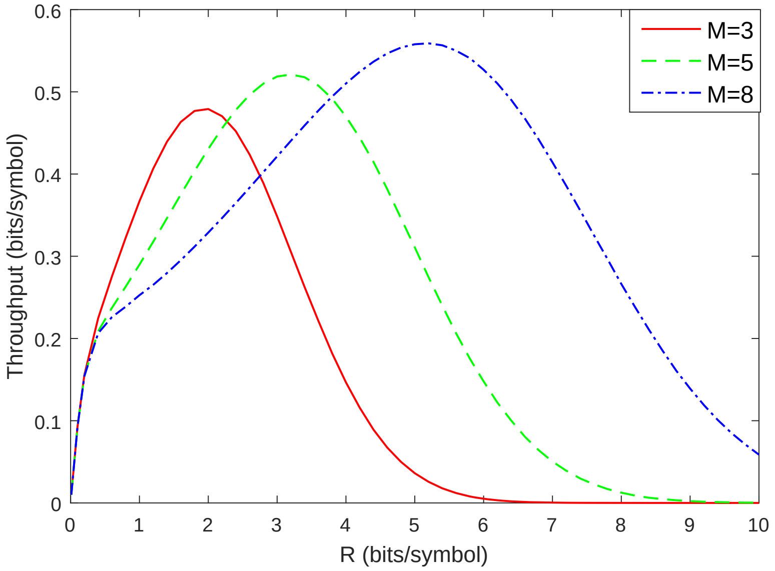

Fig. 8. The impact of fixed transmission rate R on the throughput with constant arrival sources.

deceasing in the blocklength l. This is due to the fact that as a block-fading channel model is assumed, a long l indicates a slow fading. Note that for delay sensitive network under queuing constraints, a slow fading makes a strong attenuation last for a long time, which increases the probability of buffer overflows. Hence, a longer blocklength l is expected to have a stronger influence on the throughput when the system has stricter queuing constraints. This confirms another observation from Fig. 9 that the throughput curves with a larger θ (corresponding to a stricter queuing constraints) decrease faster in l.

VII. Conclusion

We have studied in this work the FBL throughput of a low-latency IoT system operating with deadline limits, and statistical queuing constraints. Throughput characterizations have been analyzed for both the constant-rate and ON-OFF discrete time Markov arrivals. In particular, in the scenario with instantaneous CSI, we have analyzed the optimal power control maximizing the throughput, and proposed an algorithm

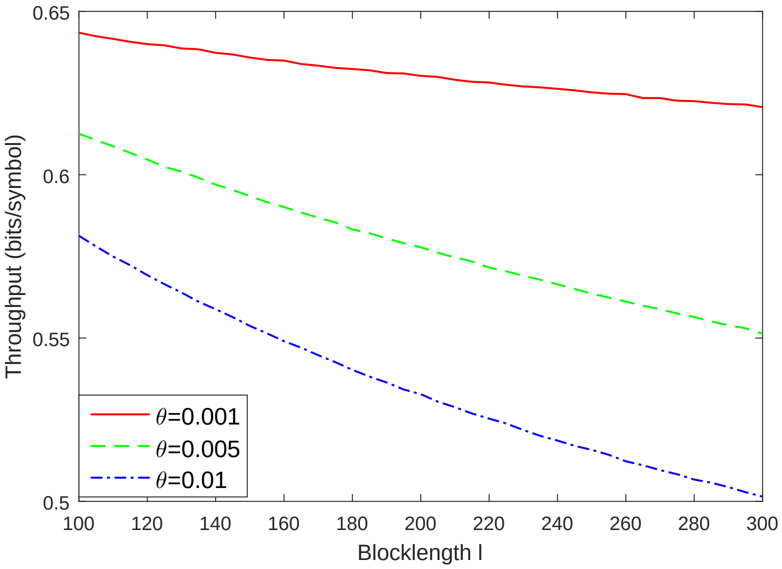

Fig. 9. The impact of blocklength l on the throughput.

to determine the optimal power levels. In addition, for the scenario with no CSI at the transmitter, we have employed HARQ-IR to improve the FBL performance. The decoding error probability and the outage probability of HARQ-IR have been characterized, following which the distribution of transmission period and throughput were derived. Finally, our characterizations have been verified via Monte Carlo simulations. Via numerical results, we have further investigated the impact of the error probability, transmission rate, QoS constraints and blocklength on the system throughput.

References

[1] Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput of HARQ-IR with finite blocklength codes and QoS constraints," *IEEE International Symposium on Information Theory (ISIT)* , Aachen, 2017, pp. 276-280.

[2] K. Campbell, J. Diffley, B. Flanagan, B. Morelli, B. O'Neil, and F. Sideco, "The 5G economy : How 5G technology will contribute to the global economy," *Technical. Report* Jan. 2017

[3] M. Maier, M. Chowdhury, B. P. Rimal and D. P. Van, "The tactile internet: vision, recent progress, and open challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138-145, May 2016.

[4] Z. Chang, Z. Zhou, S. Zhou, T. Ristaniemi and T. Chen, "Towards Serviceoriented 5G: Virtualizing the Networks for Everything-as-a-Service," *IEEE Access*, vol.6 pp.1480 - 1489, 04 Dec. 2017.

[5] C. She, C. Yang, and T. Q. S. Quek, "Radio Resource Management for Ultra-reliable and Low-latency Communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp 72-78, Jun. 2017.

[6] S. C. Lin and K. C. Chen, "Statistical QoS control of network coded multipath routing in large cognitive machine-to-machine networks," i *IEEE Internet Things J.*, vol. 3, no. 4, pp. 619-627, Aug. 2016.

[7] W. Guo, S. Zhou, Y. Chen, S. Wang, X. Chu and Z. Niu, "Simultaneous information and energy flow for IoT relay systems with crowd harvesting," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 143-149, Nov. 2016.

[8] C. Chen, J. Yan, N. Lu, Y. Wang, X. Yang and X. Guan, "Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks," *IEEE Trans. Emerg. Topics Comput*, vol. 3, no. 3, pp. 352-362, Sept. 2015.

[9] A. Sammoud, A. Kumar, M. Bayoumi, and T. Elarabi, "Real-time streaming challenges in Internet of Video Things (IoVT)," *IEEE ISCAS*, Baltimore, MD, May 2017.

[10] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. on Wireless Commun.*, vol. 2, pp. 630–643, Jul. 2003.

[11] Deli Qiao, M. C. Gursoy and S. Velipasalar, "Effective capacity region and optimal power control for fading broadcast channels," in Proc *IEEE ISIT*, St. Petersburg, 2011, pp. 2974-2978.

[12] D. Qiao, M. Ozmen and M. C. Gursoy, "QoS-driven power control in fading multiple-access channels with random arrivals," in Proc *IEEE ICC*, Kuala Lumpur, 2016, pp. 1-6.

[13] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck and M. Debbah, "Energy-efficient power control: a look at 5G wireless technologies," *IEEE Trans. on Signal Process.*, vol. 64, no. 7, pp. 1668-83, Apr.,2016.

[14] S. Efazati and P. Azmi, "Effective capacity maximization in multirelay networks with a novel cross-layer transmission framework and power-allocation scheme," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1691-1702, May 2014.

[15] W. Cheng, X. Zhang and H. Zhang, "QoS-aware power allocations for maximizing effective capacity over vrtual-MIMO wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2043-57, Oct. 2013.

[16] Q. Du and X. Zhang, "QoS-aware base-station selections for distributed MIMO links in broadband wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1123–1138, Jun. 2011.

[17] Q. Du, Y. Huang, P. Ren, and C. Zhang, “Statistical delay control and QoS-driven power allocation over two-hop wireless relay links,” in Proc *IEEE GLOBECOM*, Houston, TX, USA, Dec. 2011, pp. 1–5.

[18] Y. Hu, M. Ozmen, M. C. Gursoy and A. Schmeink, "Optimal power allocation for QoS-constrained downlink multi-user networks in the finite blocklength regime," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 9, pp. 5827-5840, Sep. 2017.

[19] X. Mi, L. Xiao, M. Zhao, X. Xu and J. Wang, "Statistical QoS-driven resource allocation and source adaptation for D2D communications underlaying OFDMA-based cellular networks," *IEEE Access*, vol. 5, pp. 3981-3999, Mar. 2017.

[20] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. on Commun.*, vol. 65, no. 4, pp. 1849-1858, April 2017.

[21] S. Wicker, *Error Control Systems for Digital Communication and Storage*. Prentice Hall, 1995.

[22] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the Gaussian collision channel,” *IEEE Trans. Inf. Theory*, vol. 47, pp. 1971–1988, Jul. 2001.

[23] P. Wu and N. Jindal, “Performance of Hybrid-ARQ in block-fading channels: A fixed outage probability analysis,” *IEEE Trans. on Commun.*, vol. 58, pp. 1129–1141, Apr. 2010.

[24] J. Choi and J. Ha, “On the energy efficiency of AMC and HARQ-IR with QoS constraints,” *IEEE Trans. Veh. Technol.*, vol. 62, pp. 3261–3270, Sep. 2013.

[25] P. Larsson, J. Gross, H. Al-Zubaidy, L. K. Rasmussen, and M. Skoglund, “Effective capacity of retransmission schemes: A recurrence relation approach,” *IEEE Trans. on Commun.*, vol. 64, pp. 4817–35, Nov. 2016.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Dispersion of gaussian channels," *IEEE ISIT*, pp. 2204–2208, IEEE, 2009.

[27] Y. Polyanskiy, H. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,”*IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, May 2010.

[28] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2430-2438, May 2015.

[29] C. She, C. Yang, and Tony Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications", *IEEE Trans. on Commun.*, vol. 66, no. 5, pp. 2266-2280, May 2018.

[30] M. Haghifam, M. Robat Mili, B. Makki, M. Nasiri-Kenari and T. Svensson, "Joint sum rate and error probability optimization: finite blocklength analysis," *IEEE Wireless Commun. Lett.* , vol. 6, no. 6, pp. 726-729, Dec. 2017.

[31] C. She, C. Yang, and Tony Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 1, pp. 127-141, Jan. 2018.

[32] M. Gharbieh, H. ElSawy, A. Bader, and M.-S. Alouini, “Spatiotemporal stochastic modeling of IoT enabled cellular networks: scalability and stability analysis," *IEEE Trans. on Commun.*, vol. 65, no. 8, pp. 3585–3600, Aug. 2017.

[33] S. Tanwir and H. Perros, “A survey of VBR video traffic models,” *IEEE Commun. Surveys Tuts.* , Jan. 2013.

[34] C.-S. Chang, “Stability, queue length, and delay of deterministic and stochastic queueing networks,” *IEEE Transactions on Automatic Control*, vol. 39, pp. 913–931, May 1994.

[35] M. Ozmen and M. C. Gursoy, "Energy-efficient power control in fading channels with Markovian sources and QoS constraints," *IEEE Trans. Commun.* , vol. 64, no. 12, pp. 5349-5364, Dec. 2016.

[36] M. Shehab, E. Dosti, H. Alves and M. Latva-aho, “Statistical QoS provisioning for MTC networks under finite blocklength”, *EURASIP J. on Wireless Commun. Netw.*, 2018:194, Aug 2018.

[37] 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, Jul. 2014.

[38] G. Ozcan and M. C. Gursoy, “Throughput of cognitive radio systems with finite blocklength codes,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541-2554, Nov. 2013.

[39] B. Makki, T. Svensson, and M. Zorzi,“Finite block-length analysis of spectrum sharing networks using rate adaptation,” *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2823–2835, Aug 2015.

[40] S. Xu *et al.*, “Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless. Commn.*, vol.15, no.8, pp.5527-5540, Aug. 2016.

[41] M. C. Gursoy, “Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP Journal on Wireless Communications and Networking*, Dec. 2013.

[42] C. Chang, *Performance Guarantees in Communication Networks*. Springer London, 2000.

[43] Y. Li, G. Ozcan, M. C. Gursoy, and S. Velipasalar, “Energy efficiency of hybrid-ARQ under statistical queuing constraints,” *IEEE Trans. Commun.*, vol. 64, pp. 4253–4267, Oct. 2016.

[44] Y. Li, M. Gursoy, and S. Velipasalar, “On the throughput of hybrid-ARQ under statistical queuing constraints,” *IEEE Trans. Veh. Technol.*, vol. 64, pp. 2725–2732, Jun. 2015.