Effective Energy Efficiency of Ultra-reliable Low Latency Communication

Mohammad Shehab, Hirley Alves, Eduard A. Jorswieck, Endrit Dosti, and Matti Latva-aho

***Abstract*—Effective Capacity defines the maximum communication rate subject to a specific delay constraint, while effective energy efficiency (EEE) indicates the ratio between effective capacity and power consumption. We analyze the EEE of ultra-reliable networks operating in the finite blocklength regime. We obtain a closed form approximation for the EEE in quasi-static Nakagami-m (and Rayleigh as sub-case) fading channels as a function of power, error probability, and latency. Furthermore, we characterize the QoS constrained EEE maximization problem for different power consumption models, which shows a significant difference between finite and infinite blocklength coding with respect to EEE and optimal power allocation strategy. As asserted in the literature, achieving ultra-reliability using one transmission consumes huge amount of power, which is not applicable for energy limited IoT devices. In this context, accounting for empty buffer probability in machine type communication (MTC) and extending the maximum delay tolerance jointly enhances the EEE and allows for adaptive retransmission of faulty packets. Our analysis reveals that obtaining the optimum error probability for each transmission by minimizing the non-empty buffer probability approaches EEE optimality, while being analytically tractable via Dinkelbach's algorithm. Furthermore, the results illustrate the power saving and the significant EEE gain attained by applying adaptive retransmission protocols, while sacrificing a limited increase in latency.**

***Index Terms*—Effective energy efficiency, finite blocklength, URLLC, IoT, optimal power allocation.**

I. Introduction

The new generations of mobile communication are expected to support a multitude of smart devices interconnected via machine type networks, enabling the Internet of Things (IoT). Energy efficient transmission while guaranteeing quality-of-service (QoS) is an ultimate goal in the design of future wireless networks. QoS constraints ranging from low latency in the order of few milliseconds and packet loss rate ($< 10^{-4}$) are key requirements for *Ultra-Reliable Low Latency Communication* (URLLC) [1]. In order to boost throughput and reliability while guaranteeing low latency, it becomes crucial to investigate and optimize the resources that are allocated for transmission. In most cases, URLLC devices such remote sensors have limited energy resources which dictates careful planning of throughput maximization with wise energy consumption models [2]. Furthermore, besides possible health risks of electromagnetic radiation in over populated areas such as city centers [3], the information and communication technology industry is projected to contribute to 6% of global CO_2 emission in 2020 [4]. This urges the invention of low power consumption, green communication schemes, yet able to perform with QoS guarantees.

In order to satisfy extremely low latency in real time applications and emerging technologies such as e-health, industrial IoT, and autonomous vehicles, an attractive solution is communication with short messages [5]. Therein, the lengths of the packets to be communicated are short, but their importance is extremely high. When the packets are short and delay requirements are stringent, performance metrics, such as Shannon capacity or outage capacity, provide a poor benchmark, and therefore, fundamentally new approaches are needed [6], [7]. In this context, the maximum achievable rate of finite blocklength packets was defined in [8] as a function of blocklength and error probability.

As envisaged by [9], the design of URLLC focuses on the tail distributions of reliability and latency instead of average metrics. Here arises the challenge of how to incorporate energy efficiency with the data rates, delay, and reliability requirements imposed by the International Telecommunication Union (ITU) and for MTC towards 6G. In this sense, metrics such as effective capacity (EC) and effective energy efficiency (EEE) are meant to capture tail statistical delay requirements in parallel with transmission throughput.

A. Related Work

The effective capacity metric was first introduced in [10] to guarantee statistical QoS requirements by capturing the physical and link layers aspects. It maps the maximum arrival rate that can be supported by a network with a maximum delay bound of δ and a delay outage probability. Unlike the the Delay-Sensitive Area Spectral Efficiency metric which only accounts for the transmission delay [11], the EC accounts for the statistical QoS aspect in terms of delay outage probability and the maximum delay bound. In [6], Gursoy characterized the EC in bits per channel use (bpcu) for short packets in quasi-static fading channels where the channel coefficients remain constant for the whole time spanning one packet transmission. Moreover, in [12], Gursoy et al. extended their analysis to multiple users but without considering the power consumption

M. Shehab, H. Alves, Endrit Dosti, and M. Latva-aho are with Centre for Wireless Communications (CWC), University of Oulu, Finland. Email: firstname.lastname@oulu.fi. Eduard A. Jorswieck is with the Department of Information Theory and Communication Systems, Technische UniverstÃt Braunschweig, Germany, Email: e.jorswieck@tu-bs.de. E. Dosti is with the Department of Signal Processing and Acoustics, Aalto University, Finland. Email: endrit.dosti@aalto.fi

This work is partially supported by Academy of Finland 6Genesis Flagship (Grant no. 318927), Aka Project EE-IoT (Grant no. 319008). The work of E. Jorswieck is partly funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy - EXC 2050/1 - Project ID 390696704 - Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

and energy efficiency aspects. Meanwhile, the per-node EC in massive MTC networks was studied in [13] proposing three methods to alleviate interference namely power control, graceful degradation of delay constraint and the hybrid method.

Effective energy efficiency is defined as the ratio between effective capacity and the total power consumption. In this sense, EEE captures the interplay between power, delay, and reliability and thus, fits well for dealing with the inherent energy-limited and bursty traffic scenarios in MTC characterized by URLLC. The maximization of EEE is of great importance for green IoT, where the goal is to maximize the throughput for each consumed unit of power. The EEE can be used as a measure of how efficient an IoT system is in terms of power consumption and energy saving. This is a very useful metric in the study and design of remote IoT devices that run on installed batteries with limited energy supply and are required to communicate with URLLC requirements. We refer to the novel factors that affect energy efficiency in MTC, which are strict delay and error constraints, bursty traffic, empty buffer probability, and communication on finite blocklength packets. In [14], the empty buffer probability (EBP) model was considered as an EEE booster for long packets transmission. The energy efficiency gap which results from utilizing finite blocklength packets and the optimum power allocation in this case were characterized in [15] for Rayleigh fading channel. The trade-off between EEE and EC was studied in [16], where the authors suggested an algorithm to maximize the EC subject to EEE constraint. However, the probability of transmission error that appears in finite blocklength communication due to imperfect coding was not considered. The authors of [17] showed that the relation between EEE and delay in wireless systems is not always a trade-off. They concluded that a linear relation between service rate and power consumption leads to an EEE-delay non-trade-off region.

On the other hand, it has been well-established that achieving ultra reliability requires the utilization of diversity schemes such as ARQ retransmission protocols. The utilization of this family of protocols has been embraced by several up to date systems such as 5G NR [18]. In this context, the authors of [19] discussed the EC of ARQ schemes for matrix-exponential distributed fading channels. They suggested that exploiting spatial diversity as the case in MIMO would reduce the sensitivity of EC to variations in the delay exponent θ. However, only the work in [20] studied the finite blocklength effect in ARQ Assisted URLLLC but without accounting power consumption as in the EEE metric.

B. Contributions

In this work, we build upon [15] and propose a finite blocklength model for the EEE in quasi-static Nakagami-m fading assuming a linear power consumption model. We characterize the optimum power allocation strategy that maximizes the EEE. We account for the EBP and in power consumption model and prove that this model is valid for short packets. Our analysis indicates that considering EBP with short packets allows for a more precise characterization of the EC and EEE. Results show that higher EC and EEE are obtained under EBP in contrast to the full buffer scenario. Besides, we illustrate the EEE gap between infinite and the finite blocklength models, and highlight the effect of network congestion on the EEE performance. We analyse how the optimum power allocation is affected by limiting the packet length and the performance gap that appears accordingly for different types of fading. In addition, we evaluate the trade-off between the EEE and the delay tolerance.

A key contribution of this work is that we exploit the non-EBP model by incorporating retransmission of faulty packets in the instants when the buffer is empty, which renders an EBP-ARQ scheme. We evaluate the EEE of the proposed scheme and also compare this case to the basic EBP model. The results show that this scenario grants a significant improvement in the EEE and reduction in the power consumption. Our analysis characterizes the optimum transmission error probability for each transmission so that the global EEE is maximized and the power consumption is minimized. The average latency is also analyzed for this scenario where the results indicate a very limited increase in latency as a cost of the high gain in EEE.

The contributions of this work are summarized as follows:

- We derive closed form approximation for the EEE under quasi-static fading.
- We characterize the optimum power allocation that maximizes the EEE[1] for linear and EBP power consumption models and highlight the trade-off between EEE and latency.
- We show the performance and optimal power allocation gap that results from utilizing short packets when compared to finite blocklength. Unlike [21], where we maximize effective capacity via optimal power allocation under Rayleigh fading while neglecting energy consumption of the devices, herein we analyze the optimal power allocation for different Nakagami-m fading setups.
- We present a basic framework for applying ARQ by considering retransmission of faulty packets in the EBP model. A buffer aware strategy is adopted to allow for rate adaption when the buffer is empty in the time slot following the current transmission. The proposed framework allows for a significant rate gain and renders a significant boost in the EEE, while maintaining a limited increase in average latency.
- We solve the optimization problem for optimum error allocation at each transmission with a target reliability constraint via low complexity Dinkelbach's algorithm. The solution shows that minimizing the non-EBP jointly reduces the power consumption and maximizes the EEE.

C. Outline

The rest of the paper is organized as follows: in Section II, we introduce the system model and clarify the relation between EC and EEE. Next, Section III presents the EEE analysis in Nakagami-m quasi-static fading and characterizes optimum

[1]It is worth noticing that in [15] we evaluated only numerically the optimal power allocation that maximizes the EEE under Rayleigh fading and without any assumptions on the EBP.

error and power allocation for the linear power consumption model. We illustrate the EBP model in finite blocklength transmission and characterize the EEE maximization with reliability, latency, and power consumption constraints in Section IV. After that, Section V studies the retransmission of faulty packets in the empty buffer instants, while the results are discussed in Section VI. Finally, Section VII concludes the paper. To make the paper more tractable, we summarize the key abbreviations and symbols that will appear throughout the paper in Table I.

TABLE I
IMPORTANT ABBREVIATIONS AND SYMBOLS.

bpcu	bits per channel use
max	maximize
NBP	non-empty buffer probability
s.t	subject to
m	fading parameter
n	blocklength
P_t	power consumption
p_{nb}	non-empty buffer probability
r	normalized achievable rate
C_e	effective capacity
δ	maximum delay
$\mathbb{E}[\ ]$	expectation of
Λ	delay outage probability
$Q(x)$	Gaussian Q-function
θ	delay exponent
ϵ_t	target error probability
ϵ^*	optimum error probability
η_{ee}	effective energy efficiency
λ	arrival rate
ρ	average signal-to-noise ratio
ζ	inverse drain efficiency

II. SYSTEM MODEL AND PRELIMINARIES

We consider a communication scenario in which an energy-limited sensor transmits data to a common aggregator through a quasi-static Nakagami-m fading channel. The received vector $\mathbf{y} \in \mathbb{C}^n$ is

$$\mathbf{y} = h\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^n$ is the transmitted packet, and the block flat-fading coefficient is denoted by $h \in \mathbb{C}$ which is assumed to be independent and identically distributed (i.i.d). This implies that h remains constant over the blocklength n, but changes from block to block. The blocklength is assumed to be smaller than the channel's coherence time. Lastly, $\mathbf{w}$ is the additive complex Gaussian noise vector whose entries are of unit variance. We assume that CSI is available at each node. Note that CSI acquisition at the transmitter in the URLLC setup is feasible whenever the channel state remains constant over multiple symbols[2]. In most communication environments, the channel coherence time is much larger than the URLLC transmissions in mini-slots of duration 0.1 ms, and thus spans of multiple TTI. This gives the transmitter sufficient time to perfectly acquire CSI [22]. Recent machine learning tools facilitate this task specially if the channel coefficients are highly correlated within a short period of time. In this case, the transmitter node can exploit a recently received signal from the other node or request a training sequence in order to estimate the channel [23]. Additionally, as in [5], we aim to provide a performance benchmark for energy efficiency of these networks, where the effect of imperfect CSI is beyond the scope of our work.

A. Communication at Finite Blocklength

In finite blocklength transmission, unlike Shannon's model, short packets are conveyed at rate that depends on the blocklength n and the packet error probability $\epsilon \in [0, 1]$, which is small but not vanishing. The normalized achievable rate, in (bpcu), is [6]

$$r(\rho) \approx \log_2(1+\rho|h|^2) - \frac{Q^{-1}(\epsilon)\log_2(e)}{\sqrt{n}}\sqrt{1-\frac{1}{\left(1+\rho|h|^2\right)^2}}, \tag{2}$$

where $Q(\cdot) = \int_{\cdot}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} \mathrm{d}t$ is the Gaussian Q-function, and $Q^{-1}(\cdot)$ represents its inverse, ρ is the average SNR which frankly represents the transmit power in watts since the noise is assumed to be normalized to unity. and $|h|^2$ is the envelope of the channel coefficients. The fading coefficients are presented by the random variable $Z = |h|^2$, which is gamma distributed with probability density function given as [24, Eq. (2.21)], [25]

$$f_Z(z) = \frac{m^m z^{m-1}}{\Gamma(m)} e^{-mz}, \tag{3}$$

where low values of m mark severe fading, high values of m mark the presence of line of sight (LOS) and $m = 1$ represents Rayleigh fading.

B. The relation between Effective Capacity and Effective Energy Efficiency

For low latency communication, effective capacity (C_e) is a powerful metric that characterizes the relation between the communication rate and the tail distribution of the packet delay violation probability [9]. For relatively large delay of multiple symbol periods, packet delay violation occurs when a packet delay exceeds a maximum delay bound δ and the outage probability is defined as [10]

$$\Lambda = \Pr(delay \geqslant \delta) \approx e^{-\theta \cdot C_e \cdot \delta}, \tag{4}$$

where $\Pr(\cdot)$ denotes the probability of a certain event. Conventionally, the tolerance of a system to long delays is measured

[2]Channel estimation is a cumbersome task for conventional systems and also under URLLC constraints. However, estimates can be reliably acquired via feedback channel in FDD system, or by exploiting channel reciprocity in TDD in line with channel inversion power control methods. CSI is obtained at reception if the latency constraint allows additional overhead due to pilot based transmissions, or alternatively via non-coherent transmissions [1].

by the delay exponent θ. The system tolerates large delays for small values of θ (i.e., $\theta \to 0$), and it becomes stricter delay-wise for large values of θ. As an exemplary scenario, consider 5G NR numerology 1 with symbol period of 35.7 micro-seconds, effective capacity of 1 bpcu and $\theta = 0.01$. For a delay outage probability of $\Lambda = 10^{-5}$ (i.e, 99.999% reliability, the network can tolerate a maximum delay of $\delta = 1151$ symbol periods (≈ 41 ms) for $\theta = 0.01$, and $\delta = 115$ symbol periods (≈ 4.1 ms) when $\theta = 0.1$. From [6], the EC in bpcu is

$$C_e(\rho, \theta, \epsilon) = -\frac{\ln\ \psi(\rho, \theta, \epsilon)}{n\theta}, \tag{5}$$

where

$$\psi(\rho, \theta, \epsilon) = \mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\rho)}\right], \tag{6}$$

and ln is the natural logarithm, $r(\rho)$ comes from (2) and $\mathbb{E}_Z[\cdot]$ denotes expectation over the fading. The above equations assumes an underlying simple ARQ process and indicate that higher service rates reduce the amount of data bits stored in the queue, and hence also reduce the delay required to transmit, which boosts the EC. Thus, the EC is a measure of the throughput while statistically guaranteeing the delay tolerance.

Remark 1. *Note that, we resort to the more practical concept of service rate as in [6] rather than the departure rate model in [10]. The intuition is that the service rate metric is more suitable for the characterization of the re-invented effective capacity (or effective rate) for discrete short packets operating in the finite blocklength regime. Herein, we consider short packets with a packet drop probability of ϵ to measure reliability. Reliability could be well mapped via service rate rather than departure rate which does not account for the dropped packets with probability ϵ. This model allows us to combine the latency and reliability aspects and accommodate them in the characterization of QoS constrained energy efficiency.*

In [6], the effective capacity is studied for single node scenario, but never to a closed form expression. It has been proven that the EC is concave in ϵ, and hence has a unique global optimum. From (2) and (5), we observe that increasing the transmission power would definitely raise the EC. However, this comes at the expense of increased power consumption, which is not suitable for energy-limited (battery-operated) IoT devices. Thus, it is necessary to study the trade-off between enhancing EC and power consumption.

In this context, effective energy efficiency is defined as the achieved effective capacity per unit of consumed power. The EEE captures the trade-off between the throughput of the communication link, the overall power consumption, and latency. Thus, EEE is a suitable metric to quantify and optimize the throughput of the communication link per each consumed watt for energy-limited, low latency IoT. Hence, the optimization of EEE is of great importance for IoT devices, which are isolated from stationary power sources and are required to deliver packets with low latency in the order of milli-seconds [26]. In what follows, we study the EEE for short packet communication.

III. Effective Energy Efficiency under linear power consumption model

In our analysis, we resort to a linear power consumption model defined as [27]

$$P_t(\rho) = \zeta\rho + P_c, \tag{7}$$

with $\zeta \geqslant 1$ being the inverse drain efficiency of the transmit amplifier and P_c the hardware power dissipated in circuit in watts. The linear power consumption model is a well-established and accepted model that has been widely used in many studies related to energy efficiency of wireless systems such as [14], [27]–[29]. This model captures the linear increase of the power consumption as a function of the transmit power and the inverse drain efficiency as well as the idle circuit power consumption. Furthermore, it facilitates the analysis that aims at providing a performance benchmark for the EEE in the context of short packet and low latency communication of IoT devices.

For this model, the EEE is given by

$$\eta_{ee} = \frac{-\frac{1}{n\theta}\ln\left(\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r}\right]\right)}{\zeta\rho + P_c}. \tag{8}$$

This scenario assumes an always full buffer and does not account for EBP. In [6], a stochastic model for EC was studied, but never to a closed form expression. Herein, we present a tight approximation for the EC and hence, the EEE.

Lemma 1. *The effective capacity in Nakagami-m quasi-static fading is approximated by*

$$C_e(\rho, \theta, \epsilon) \approx -\frac{1}{n\theta}\ln\left[\epsilon + (1-\epsilon)\frac{m^m}{\Gamma(m)} \cdot \sum_{n=0}^{2}\frac{\beta^n}{n!}\int_0^\infty (1+\rho z)^\alpha \gamma^n z^{m-1}e^{-mz}dz\right], \tag{9}$$

where $\alpha = \frac{-\theta n}{\ln 2}$, $\beta = \theta\sqrt{n}Q^{-1}(\epsilon)\log_2 e$, *and* $\gamma = \sqrt{1 - \frac{1}{(1+\rho z)^2}}$.

Proof. Please refer to Appendix A □

Remark 2. *It is quite straightforward to conclude that EEE is an increasing function of the fading parameter m. That is, the EEE becomes worse when the fading becomes more severe (i.e, $m \to \frac{1}{2}$). Hence, starting from here, we focus our analysis on quasi-static Rayleigh fading to provide a benchmark of the EC and EEE in the proposed scenarios, where the results can also be extended to any type of Nakagami-m fading. However, Lemma 1 facilitates the following derivation of the EEE in Rayleigh fading and later, comparing the optimum power allocation in each fading scenario.*

Theorem 1. *For a Rayleigh quasi-static fading channel with blocklength n, the EEE of the linear power consumption model is approximated as*

$$\eta_{ee}(\rho, \theta, \epsilon) \approx -\frac{\ln\left[\epsilon + (1-\epsilon)\ \mathcal{J}\right]}{n\theta\left(\zeta\rho + P_c\right)}, \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3052965, IEEE Internet of Things Journal

where

$$\mathcal{J}=e^{\frac{1}{\rho}}\rho^{\alpha}\left[\left(\frac{\beta^2}{2}+\beta+1\right)\Gamma\left(\alpha+1,\frac{1}{\rho}\right)-\left(\frac{\beta^2}{2}+\beta\right)\frac{\Gamma\left(\alpha-1,\frac{1}{\rho}\right)}{\rho^2}\right], \quad (11)$$

where $\Gamma(\cdot,\cdot)$ *is the upper incomplete gamma function [30].*

Proof. Please refer to Appendix B. □

Remark 3. *It was shown in [15] that both EC and EEE are concave functions of the error probability* ϵ*, and the optimum value of* ϵ *that maximizes them in this case is given by*

$$\epsilon^*(\rho,\alpha,\beta) \approx \arg\min_{0\leqslant\epsilon\leqslant 1} \ \epsilon+(1-\epsilon)\ \mathcal{J}. \quad (12)$$

In what follows, we study the behaviour of EEE as a function of transmit SNR ρ.

Theorem 2. *The EEE function in Theorem 1 is a quasi-concave function of the transmit SNR.*

Proof. Please refer to Appendix C □

Fig. 1 illustrates the EEE in Rayleigh block fading channel for different delay exponents, while applying the expectation in (8) and Theorem 1. The network parameters are $n = 500$ symbol periods and $\epsilon = 10^{-4}$. The figure proves the accuracy of Theorem 1 as a tight approximation for the EEE, which is also well established for the EC in [21]. However and unlike the EC which is an asymptotically increasing function of the SNR, the figure shows the convexity of the upper contour of the EEE in the transmit power and that the approximation in Theorem 1 captures this quasi-concavity precisely as stated in Theorem 2. Note that the EEE declines when the delay constraint becomes more strict. Meanwhile, it is also observed that the optimum transmit power shifts to a higher value for less strict delay constraints. Finally, we plot the EEE also for smaller packets with length of $n = 50$ symbol periods. As we can observe, the EEE is higher for smaller packet size as the delay is minimized which boosts the EC. The approximation holds tightly for this setup as well.

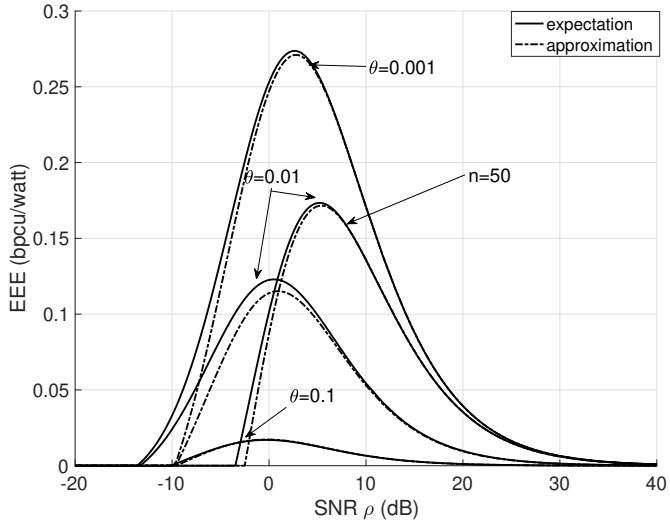

Fig. 1. EEE vs SNR in quasi-static fading for $m = 1$, $n = 500$, $\epsilon = 10^{-4}$, $P_c = 1.2$, $\zeta = 1.2$, $\lambda = 1$ and different delay exponents θ.

The quasi-concavity of the EEE in the SNR aids to characterize the optimum power allocation which maximizes the EEE in the linear power consumption model.

Theorem 3. *The optimum power allocation for maximizing the EEE is* ρ^**, which is the solution of*

$$\eta_{ee}(\rho^*) = -\frac{1}{n\theta}\left(\frac{\mathcal{J}'(\rho^*)}{\mathcal{J}(\rho^*)}\right), \quad (13)$$

where $\kappa_1 = \frac{\beta^2}{2}+\beta+1$ *and* $\kappa_2 = \frac{\beta^2}{2}+\beta$*, and*

$$\mathcal{J} = e^{\frac{1}{\rho}}\rho^{\alpha}\left(\kappa_1\Gamma\left(\alpha+1,\frac{1}{\rho}\right)-\frac{\kappa_2}{\rho^2}\Gamma\left(\alpha-1,\frac{1}{\rho}\right)\right), \quad (14)$$

$$\begin{aligned}\mathcal{J}'(\rho^*) &= \frac{\partial\mathcal{J}}{\partial\rho}\\ &= -\frac{1}{\rho^2}\left[\left(1+\frac{\alpha}{\rho}\right)\mathcal{J}+\frac{(1-\kappa_1)e^{-\frac{1}{\rho}}}{\rho^{\alpha}}-\frac{2\kappa_2}{\rho}\Gamma\left(\alpha-1,\frac{1}{\rho}\right)\right].\end{aligned} \quad (15)$$

Proof. Based on the quasi-concavity of the EEE function which was proven in Theorem 2, we differentiate (10) with respect to ρ and equate to zero as follows

$$\begin{aligned}\frac{\partial\eta_{ee}}{\partial\rho} &= -\left[\frac{\frac{(1-\epsilon)\mathcal{J}'(\zeta\rho+P_c)}{\epsilon+(1-\epsilon)\mathcal{J}}-\zeta\ln(\epsilon+(1-\epsilon)\mathcal{J})}{n\theta(\zeta\rho+P_c)^2}\right]\\ &\approx -\left[\frac{\frac{\mathcal{J}'(\zeta\rho+P_c)}{n\theta\mathcal{J}(\zeta\rho+P_c)}-\frac{\zeta\ln(\epsilon+(1-\epsilon)\mathcal{J})}{n\theta(\zeta\rho+P_c)}}{(\zeta\rho+P_c)}\right]=0,\end{aligned} \quad (16)$$

where the above approximation is valid since $\mathcal{J}$ is much larger than 1, and the reliability constraint ϵ is very small (i.e, $\mathcal{J} >> \epsilon$). Then, to differentiate $\mathcal{J}$, we apply the derivative of the upper incomplete gamma function [31], which yields

$$\frac{\partial\mathcal{J}}{\partial\rho}=-\frac{1}{\rho^2}\left[\mathcal{J}+\frac{\alpha}{\rho}\mathcal{J}-\frac{\kappa_1 e^{-\frac{1}{\rho}}}{\rho^{\alpha}}-\frac{2\kappa_2}{\rho}\Gamma\left(\alpha-1,\frac{1}{\rho}\right)+\frac{e^{-\frac{1}{\rho}}}{\rho^{\alpha}}\right], \quad (17)$$

and after algebraic manipulation we obtain (15), which concludes the proof. □

Despite the fact that we were able to find the partial derivative of $\mathcal{J}$, a closed form solution for (13) does not exist. For this purpose, we can compute a point-wise numerical solution or utilize Matlab root-finding functions, e.g., fzero in a similar way to [16].

IV. Empty buffer probability model

Previously, we assumed that the buffer is always full which practically is not always the case. In real scenarios, there would be instants in which a certain IoT device becomes idle and therefore has no data to transmit. Thus, we need to account for the case when the buffer is empty. Accordingly, we apply the model considered in [14] to networks operating in the finite blocklength regime with non-vanishing probability of error ϵ. After accounting for EBP, the transmission probability P_{nb} is equal to (1− the probability of empty buffer) and the transmission process appears in Fig. 2.

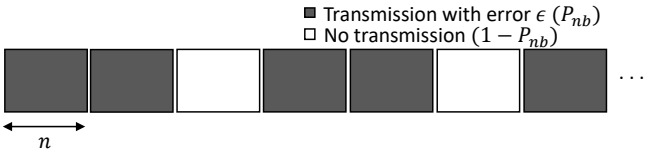

Fig. 2. Transmission with empty buffer probability in quasi-static channel with blocklength n.

For an average arrival rate of λ and a stable queue, the power consumption becomes

$$P_t(\rho) = P_{nb}\zeta p + P_c = \frac{\lambda}{\mathbb{E}[r]}\zeta\rho + P_c, \tag{18}$$

with $P_{nb} = \frac{\lambda}{\mathbb{E}[r]}$ denoting non-empty buffer probability (NBP), which is bounded between 0 and 1. The EEE with EBP is

$$\eta_{ee} = \frac{-\frac{1}{n\theta}\ln\left[\epsilon + (1-\epsilon)\ \mathcal{J}\right]}{\frac{\lambda}{\mathbb{E}[r]}\zeta\rho + P_c}, \tag{19}$$

where the numerator represents the effective capacity in the finite blocklength regime as defined in Theorem 1.

Note that similar to the fluid model[3], the NBP indicates the average asymptotic probability of transmission over relatively long time, where $P_{nb} = 1$ means that the average serviced amount of data per packet time is equal to the average amount of data arrival per packet time. This indicates that, in average, transmission always occurs.

A. Verifying the effective energy efficiency model with empty buffer probability in finite blocklength

According to [28], an energy efficiency function must be non-negative, must be zero when the transmit power is zero, and must tend to zero as the transmit power tends to infinity. It was shown in [14] that this power the EBP model fulfills is valid for Shannon model. In the following Lemma, we verify that this EBP power consumption model is valid as well for short packets transmission.

Lemma 2. *The EEE in (19) is zero for $\rho = 0$ and tends to 0 when $\rho \to \infty$.*

Proof. Please refer to Appendix D. □

B. Effective energy efficiency maximization with buffer constraints

We investigate the EEE maximization with EC, delay, and power constraints. EC should be larger than the arrival rate λ to guarantee a stable queue, while the transmission SNR ρ is bounded by ρ_{max}. Thus, the optimization problem is formulated as

$$\begin{aligned} \max_{\rho\geqslant 0,\theta\geqslant 0}\ & \eta_{ee} = \frac{-\frac{1}{n\theta}\ln\left[\epsilon + (1-\epsilon)\ \mathcal{J}\right]}{P_{nb}\zeta\rho + P_c}, \\ s.t\ \ & C_e(\rho,\theta,\epsilon) \geqslant \lambda, \quad P_{nb}e^{-\theta\lambda\delta} \leqslant \Lambda \\ & \rho \leqslant \rho_{max}, \quad \epsilon \leqslant \epsilon_t, \quad 0 \leqslant P_{nb} \leqslant 1. \end{aligned} \tag{20}$$

Note that the reliability constraint here, which is the error probability constraint on the first transmission, is important to improve QoS in URLLC. Meanwhile, the delay outage probability constraint does not necessarily guarantee reliability if the ϵ is not imposed.

For the full buffer model, we set P_{nb} to 1. We perform a line search for ρ in the interval $[0, \rho_{max}]$. The optimum error probability is $\min[\epsilon^*, \epsilon_t]$ where ϵ^* is obtained from Remark 3. When analyzing the empty buffer scenario, we set $P_{nb} = \frac{\lambda}{\mathbb{E}[r]}$. Here, Λ is the maximum allowed delay outage probability. In all cases, the optimal value of θ can be obtained from the second constraint at equality as

$$\theta^*(\rho) = \frac{1}{\lambda\delta}\ln\frac{P_{nb}}{\Lambda}. \tag{21}$$

V. Adaptive Retransmission scenario

As shown in [32], achieving ultra-reliability using one transmission consumes a huge amount of power, which is not applicable for energy limited IoT devices. Herein, we present a basic framework for applying ARQ by considering adaptive rate retransmission of faulty packets when the buffer is empty in the EBP model in order to achieve ultra-reliability. In this framework, at a certain time instant t, we assume a buffer-aware transmission as in [33], [34]; this allows the transmitter to have a prior knowledge of whether the buffer will be empty or there is a packet that needs to be delivered at the next time slot $t+1$. Then the following is applied:

1) If a packet arrives at time slot t and there is also a packet arrival at $t+1$, normal transmission occurs at t with rate $r(\epsilon)$.
2) If a packet arrives at time slot t and there is no packet arrival at $t+1$, we transmit the with rate $r(\epsilon_1 > \epsilon)$ and apply ARQ at $t+1$.

However, the non-EBP in this case will be slightly changed to $P_{nb}^{'}$ due to the rate variations. The first case occurs when the buffer is full at $t+1$. Assuming i.i.d arrivals, the probability of occurrence of case number 1 is the same as the non-empty buffer probability $P_{nb}^{'}$.

In case 2, we apply the type-I ARQ protocol. In type-I ARQ protocol, the node is allowed to retransmit its packet if it receives a NACK feedback from the receiver, which indicates that the packet is not successfully decoded. We assume a maximum of only 1 retransmission in order to satisfy the stringent delay requirements in URLLC. Note that, the transmitter performs the first transmission with an error probability of ϵ_1 and the second transmission with an error probability of ϵ_2 such that the aggregate error probability satisfies the reliability constraint (i.e, $\epsilon_1\epsilon_2 \leqslant \epsilon$). Thus, both ϵ_1 and ϵ_2 are higher than ϵ. The fact that the transmission rate is an increasing function of the error probability implies that both $r(\epsilon_1)$ and $r(\epsilon_2)$ are

[3]The reason behind the choice a fluid model is that these fluid models are motivated as approximations to discrete queueing systems. Fluid flow queues have been well accepted as a useful mathematical tool for modeling and have long been used to evaluate the performance of telecommunication and computer systems. In particular, we apply the fluid model to characterize the asymptotic delay probability, and to approximate the NBP by arrival rate divided by average service rate. The exact (non-asymptotic) analysis of delay outages and empty buffers is outside the scope of this work.

higher than $r(\epsilon)$, which reflects the rate gains of this model. In other words, we obtain a significant rate gain in the likely event of successful first transmission. Moreover, the aggregate packet error probability when applying ARQ would be ϵ which maintains the reliability level. This can be performed by rate adaption as in [35], where sensors generate data in the form of bits that can be relocated from one packet to the other. This rate adjustment occurs when the buffer is empty in the next time instant $t+1$ with probability $1-P_{nb}^{'}$ and results in a rise in the instantaneous rate with symbols resizing. Generally let $r_0 = \mathbb{E}[r(\epsilon)]$, $r_1 = \mathbb{E}[r(\epsilon_1)]$ and $r_2 = \mathbb{E}[r(\epsilon_2)]$. Then the average rate becomes

$$\begin{aligned}\mathbb{E}[r] &= p_{nb}^{'} r_0 + (1-p_{nb}^{'})\left[(1-\epsilon_1)r_1 + \frac{\epsilon_1 r_2}{2}\right] \\ &= p_{nb}^{'}(r_0-\kappa)+\kappa, \end{aligned} \tag{22}$$

where $\kappa = (1-\epsilon_1)r_1 + \frac{\epsilon_1 r_2}{2}$. Note that the rate is divided by 2 in case of retransmission because the time duration of transmitting n symbols is approximately the double as expressed in the last term of (22). Due to the rate variation, the modified non-EBP should satisfy

$$p_{nb}^{'} = \frac{\lambda}{\mathbb{E}[r]} = \frac{\lambda}{p_{nb}^{'}(r_0-\kappa)+\kappa}. \tag{23}$$

Solving (23) for $p_{nb}^{'}$, we obtain

$$p_{nb}^{'} = \frac{-\kappa + \sqrt{\kappa^2 + 4(r_o-\kappa)\lambda}}{2(r_o-\kappa)} \leqslant 1, \tag{24}$$

and the EC in this case is given by

$$\begin{aligned} C_{e2} = \frac{-1}{n\theta}\ln\Big(\mathbb{E}_Z\Big[p_{nb}^{'}(\epsilon+(1-\epsilon)e^{-n\theta r_0}) + \\ (1-p_{nb}^{'})\left((1-\epsilon_1)e^{-n\theta r_1} + \epsilon_1(1-\epsilon_2)e^{-n\theta\frac{r_2}{2}} + \epsilon\right)\Big]\Big). \end{aligned} \tag{25}$$

To clarify (25), we mention that the transmission occurs with rate r_0 when the buffer is not empty in the next time slot. This is indicated by the first term in equation (25) and occurs with probability $p_{nb}^{'}$. When the buffer is empty in the next slot with probability $(1-p_{nb}^{'})$, one transmission occurs rate r_1 if there is no error where the no error probability is $1-\epsilon_1$. However, the rate is divided by 2 to become $r_2/2$ in case of retransmission when an error occurs in the first transmission only with probability $\epsilon_1(1-\epsilon_2)$. While the rate is considered to be zero when both transmissions fail with probability $\epsilon_1\epsilon_2 = \epsilon$. In order to map the impact of the second transmission on the overall effective capacity, we calculate the average of the rate of the first transmission which is zero in case of error and the second transmission which is $r(\epsilon_2)$. Since the transmission occurs in the duration of 2 time slots, the rate is virtually divided by 2 as indicated in the second term of (25).

Since the channel coefficients change from one transmission to the other, we need to vary the transmit power in order to compensate for the channel coefficient variation so that the product $\rho|h|^2$ is the same for both transmissions. Although the transmit power varies, the average transmit power $\mathbb{E}[\rho]$ is the same for both transmissions and independent from $p_{nb}^{'}$ or any other parameter. Thus, the consumed power for this scenario is a probabilistic function of the transmit power of one or two transmissions which after manipulations is allowed to be expressed as

$$P_t = \left[p_{nb}^{'2} + p_{nb}^{'}(1-p_{nb}^{'})(1+\epsilon_1)\right]\zeta\rho + P_c, \tag{26}$$

Defined by the quotient of EC to the transmit power, the EEE of this scenario is formulated as

$$\eta_{ee_2} = \frac{C_{e2}}{\left[p_{nb}^{'2} + p_{nb}^{'}(1-p_{nb'}^{'})(1+\epsilon_1)\right]\zeta\rho + P_c}. \tag{27}$$

Note that the optimum error probability of each transmission in the EBP model with two ARQ transmissions is not simply the square root of the aggregate target error probability $\epsilon = \epsilon_t$. Therefore, we define the optimization problem to determine the optimum error probability of the first transmission that maximizes the EEE subject to a target reliability constraint as

$$\begin{aligned} \max \quad & \eta_{ee_2}(\epsilon_1,\epsilon_2) \\ s.t \quad & 0 < \epsilon_1\epsilon_2 \leqslant \epsilon_t \end{aligned} \tag{28}$$

An interesting analysis is to determine the asymptotic behaviour of the EEE as the delay constraint $\theta \to \infty$ or 0. This corresponds to extremely strict or no delay constraint, respectively. It is straight forward to conclude that the EEE tends to zero for extremely stringent delay constraint (i.e, when $\theta \to \infty$). Herein, we derive the upper bound of the EEE for relaxed latency constraint as $\theta \to 0$.

Theorem 4. *The EEE of the proposed retransmission scenario is upper bounded by*

$$\lim_{\theta\to 0}\eta_{ee_2} = \frac{p_{nb}^{'}(1-\epsilon)r_o + (1-p_{nb}^{'})\left[(1-\epsilon_1)r_1 + \epsilon_1(1-\epsilon_2)\frac{r_2}{2}\right]}{\left[p_{nb}^{'2} + p_{nb}^{'}(1-p_{nb'}^{'})(1+\epsilon_1)\right]\zeta\rho + P_c}, \tag{29}$$

and lower bounded by zero.

Proof. Please refer to Appendix E. □

Remark 4. *As the system approaches ultra-reliability (i.e, $\epsilon \to 0$), the EC of one transmission converges to r_0, while the EC of the proposed EBP-ARQ scehme converges to $p_{nb}^{'}r_o + (1-p_{nb}^{'})r_1$. Hence, the EC is raised to r_1 for $(1-p_{nb}^{'})$ portion of the time which indicates the gain in the EC of the proposed EBP retransmission scheme.*

A. Power saving

Returning to (26), which represents the average power consumption, we study the effect of varying the non-EBP $p_{nb}^{'}$ on the power consumption by obtaining the first derivative of (26) as

$$\frac{\partial P_t}{\partial p_{nb}^{'}} = \left[-2\epsilon_1 p_{nb}^{'} + (1+\epsilon_1)\right]\zeta\rho, \tag{30}$$

which is strictly non-negative for all possible values of $p_{nb}^{'}$ and ϵ_1 (i.e, $0 \leqslant p_{nb}^{'}, \epsilon_1 \leqslant 1$). Thus, the power consumption p_t is still an increasing function of the non-empty buffer probability $p_{nb}^{'}$. Hence, minimizing the non-empty buffer probability also reduces the transmit power which leads to a longer battery life for remote sensors that are located far from energy sources.

Theorem 5. *The non-EBP $p_{nb}^{'}$ in (24) is a pseudo-convex function in ϵ_1 and therefore, the minimization of $p_{nb}^{'}$ is a fractional program.*

Proof. Please refer to Appendix F. □

Furthermore, being a psuedo convex fractional program and due to the analytical intractability, it is easier to find the global minimum of $p_{nb}^{'}$ using well known optimization algorithms such as Dinkelbach's algorithm [28] to minimize $p_{nb}^{'}$ and hence, the total transmit power. Later the results section shows that this optimal solution for minimizing the transmit power also highly approaches optimality for maximizing the EEE in this case. Moreover, it is more efficient and numerically tractable[4] to minimize the transmit power instead of the EEE function given in (28). Hence, the problem becomes

$$\begin{aligned} \min \quad & p_{nb}^{'}(\epsilon_1, \epsilon_2) \\ s.t \quad & 0 < \epsilon_1 \epsilon_2 \leqslant \epsilon_t. \end{aligned} \tag{31}$$

Note that the target is to minimize the transmit power. Thus, it is straight forward to conclude that the reliability constraint is optimally achieved at equality since more power is needed to achieve lower error and higher reliability. As proven in Theorem 5, the problem in (31) is a pseudo-convex fractional program. Therefore, the global optimum exists, and can be found by utilizing the Dinkelbach's algorithm as introduced in Section 3.2 in [28]. It is a parametric algorithm of which the basic idea is to tackle a pseudo-convex problem by solving a sequence of easier problems which are guaranteed to converge to the global optimum. The minimization procedure is depicted in Algorithm 1.

Algorithm 1: Minimization of $p_{nb}^{'}$

Input : $F_{\sigma_0} > \delta > 0$; $n = 0$; $\sigma = 0$;
Output: ϵ_1^*

while $F_{\sigma_n} > \delta$ **do**
  $\epsilon_1^* = \arg\min\{-\kappa(\epsilon_1) + \sqrt{\kappa(\epsilon_1)^2 + 4(r_o - \kappa(\epsilon_1))\lambda} - \sigma_n 2(r_o - \kappa(\epsilon_1))\}$;
  $F_{\sigma_n} = -\kappa(\epsilon_1^*) + \sqrt{\kappa(\epsilon_1^*)^2 + 4(r_o - \kappa(\epsilon_1^*))\lambda} - \sigma_n 2(r_o - \kappa(\epsilon_1^*))$;
  $\sigma_{n+1} = \frac{-\kappa(\epsilon_1^*) + \sqrt{\kappa(\epsilon_1^*)^2 + 4(r_o - \kappa(\epsilon_1^*))\lambda}}{2(r_o - \kappa(\epsilon_1^*))}$;
  $n = n + 1$;
end

The intuition behind the algorithm is as follows. It starts from some arbitrary estimate of ϵ_1^* and analyzes the level sets of the original problem, which are evidently convex. Then, as the algorithm progresses, it iteratively corrects the estimate of ϵ_1^* and checks if the stopping criterion is satisfied, i.e. a δ-suboptimal solution has been obtained. If the tolerance margin has not yet been satisfied, then the algorithm continues scanning through the level sets of the function until convergence. Notice that the worst-case computational complexity of the algorithm is dominated by step 2, which can be solved using interior point methods. As a consequence, the convergence rate in the sub-problem sequence is super-linear [36], [37].

[4]Taking into account the energy consumption and computational complexity for the original online optimization algorithm, it might even turn out that the suboptimal one proposed here leads to an overall better effective energy efficiency and lower latency.

B. Average latency

Herein, we analyze the average extra packet delay induced due to retransmissions when applying the proposed EBP-ARQ with empty buffer instants. Let δ_1 be the delay per packet in the single transmission scenario and δ_2 be the total delay when two transmissions occur. Then the expected delay when applying ARQ with two retransmissions would be

$$\tau = \delta_1 \left(P_{nb}^{'} + (1 - P_{nb}^{'})(1 - \epsilon_1) \right) + \delta_2 (1 - P_{nb}^{'})\epsilon_1. \tag{32}$$

In case of error in the first transmission, the 1 bit NACK feedback could be transmitted in a span of ≈ 6 symbols according the Physical Uplink Control Channel (PUCCH) Format 1 in 5G NR [38]. Assuming the same transmission rate for the NACK packet, the extra delay due to the NACK packet would be $\Delta = \frac{6}{n}\delta_1$. This occurs during before the second transmission and is very small compared to the one packet transmission time. Thus, we can state that $\delta_2 = 2\delta_1 + \Delta = \left(2 + \frac{6}{n}\right)\delta_1$. Hence, the normalized delay with respect to one transmission time δ_1 can be written as

$$\tau_n = P_{nb}^{'} + (1 - P_{nb}^{'})(1 - \epsilon_1) + \left(2 + \frac{6}{n}\right)(1 - P_{nb}^{'})\epsilon_1, \tag{33}$$

where $\tau_n \geqslant 1$. Herein, (33) provides an indication of the QoS when applying ARQ wiht EBP for boosting the EEE. Note that we mainitain the same QoS constraints θ snd ϵ throughout the whole analysis.

VI. Results and discussion

In this section, we present numerical results to illustrate the behaviour of the EEE function and the trade off between the EEE, power allocation and latency for Shannon's model and finite blocklength in different transmission scenarios. We compare our results to the infinite blocklength case to show the performance gap that results from applying the short packet information theoretic approach which is more suitable for delay constrained analysis and compare this gap to the long packets ideal case. Firstly, Fig. 3 illustrates the EEE of short packet transmission as a function of the delay exponent θ in quasi-static Rayleigh fading for $n = 500$, $\rho = 3$ dB, error probability $\epsilon = 10^{-4}$, and different circuit powers P_c. The figure highlights the energy efficiency gap between long packet transmission which is analyzed via Shannon capacity model and the finite blocklength model. The EEE of short packets is less than the infinite blocklength Shannon's model by about 20% in this case. Moreover, the figure shows that the EEE declines when the delay exponent becomes more strict and when the consumed power in circuitry is higher.

In Fig. 4, we elucidate the EEE for different transmission probabilities (i.e, when the buffer is not empty). The figure shows the EEE gap between infinite and finite blocklength models. Again, it is noted that higher circuit power significantly deteriorates the EEE. It is observed that the EEE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3052965, IEEE Internet of Things Journal

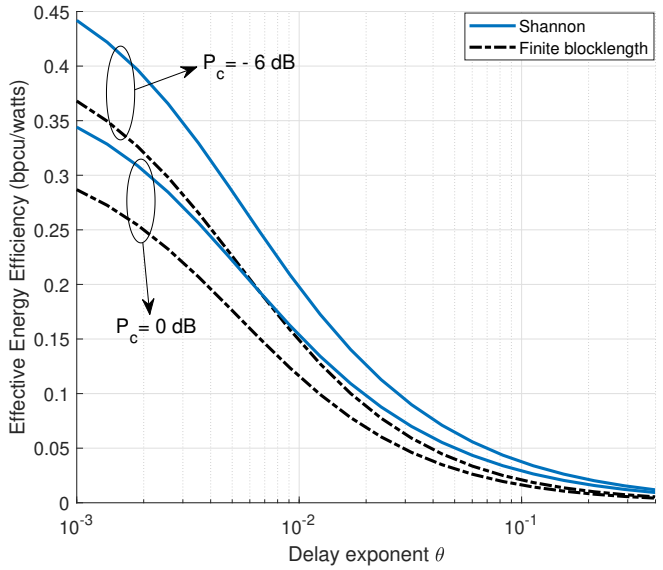

Fig. 3. Effective energy efficiency as a function of the delay exponent θ in quasi-static Rayleigh fading for $n = 500$, ρ=3 dB, error probability $\epsilon = 10^{-4}$, and different circuit powers P_c.

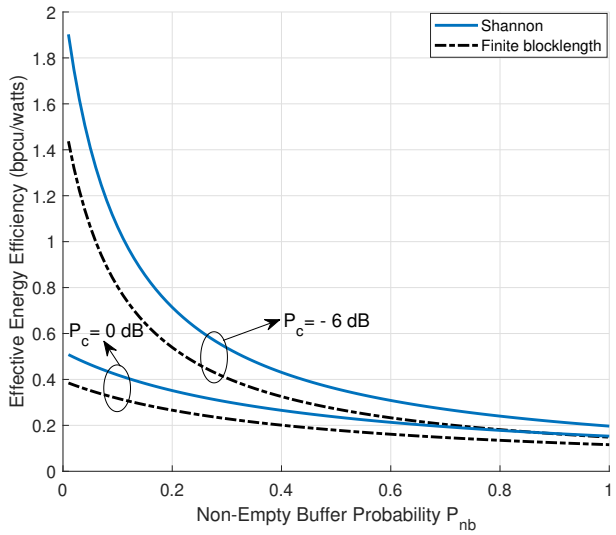

Fig. 4. Effective energy efficiency vs as a function of the non-EBP P_{nb} in quasi-static Rayleigh fading for $n = 500$, SNR=3 dB, error probability $\epsilon = 10^{-4}$, $\theta = 0.01$ and different circuit powers P_c.

monotonically decreases with the increase of arrival rate (or alternatively Non-EBP) which indicates higher congestion in the network. However, this effect becomes marginal when the circuit power is higher as the circuit power becomes a dominant factor in the calculation of EEE. Thus, for $P_c = 0$ dB, the EEE is nearly constant as a function of the arrival rate. Therefore, careful studying of EEE for different source arrival rates is crucial for low circuit power.

For the following simulations, we fix the network parameters as follows: $\Lambda = \{10^{-2}, 10^{-3}\}$, $P_c = 0.2$ W, $\zeta = 1.2$, $\lambda = 1$, $\delta = 500$ symbol periods, and $n = 500$ symbol periods, unless stated otherwise. In Fig. 5, we evaluate the EEE as a function of error probability ϵ in case of EBP and compare it to the case where the buffer is always full while fixing

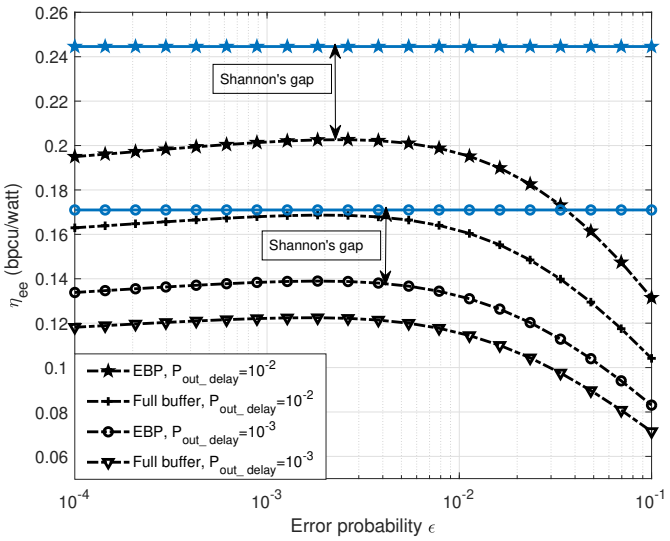

Fig. 5. EEE vs ϵ with and without empty-buffer probability for $\Lambda = 10^{-2}, 10^{-3}$, $P_c = 0.2$, $\zeta = 1.2$, $\lambda = 1$, $\delta = 500$, and $n = 500$.

the transmit power at $\rho = 10$ dB. We observe that the EEE is concave in ϵ as stated in Remark 3. It is obvious that considering the probability of empty buffer reflects a gain in the EEE over the full buffer model, while decreasing the delay outage probability reduces the EEE. Moreover, the figure depicts that Shannon's model considered in [14] overestimates the EEE by more than 20% when compared to the finite blocklength model.

Fig. 6 depicts the maximum achieved EEE obtained from (20) for different delay limits δ, where $\rho_{max} = 13$ dB (variable transmission power), and $\epsilon_t = 10^{-4}$. We observe that the EEE increases when extending the delay bound δ and relaxing the delay outage probability Λ. This implies that networks which can tolerate longer packet transmission delay are more energy efficient. From another perspective, it is clear that the sporadic non-EBP transmission scenario allows for a better modelling of the power consumption in MTC. This reflects that full buffer is the worst case, where we assume that all power will be consumed. Meanwhile, the Non-EBP models the fraction of time that is actually used for transmission of packets according to the queue congestion, which interprets the gain of this model compared to always full buffer. Furthermore, the figure verifies the inaccuracy of Shannon's model when computing the EEE for relatively small packets where the inaccuracy gap reaches more than 30% in higher delay region.

In order to present an insight about how EBP would affect the performance of multi-user network, we consider a simple exemplary setup where 2 users transmit short packets to a common BS. The BS applies successive interference cancellation where User 1 is the primary user assumed to be ultra reliable with $\epsilon_{u1} = 10^{-4}$, and therefore decoded last while it transmits with higher power 6 dB. Meanwhile, User 2 is the secondary which has lower priority, where it transmits with low transmit power of 0 dB, and reliability of $\epsilon_{u1} = 0.1$. Fig. 7 depicts that the NBP of User 1 does not only affect the EEE of User 1, but also affects both the EEE and NBP of User 2.

Taking a close look at Fig. 7, we observe that adjusting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3052965, IEEE Internet of Things Journal

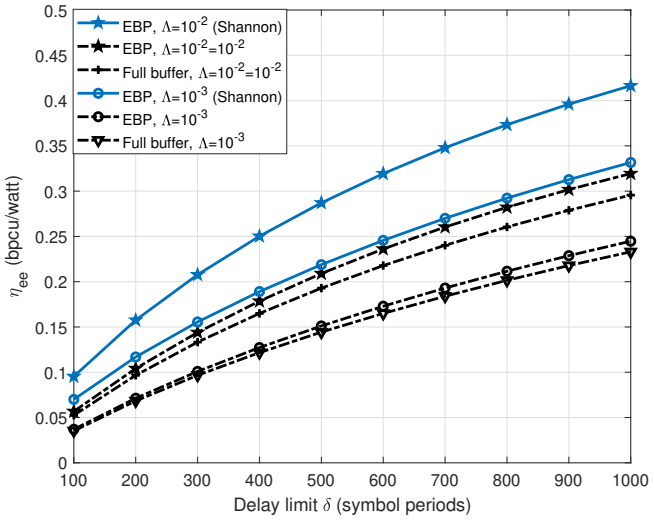

Fig. 6. EEE vs δ with and without empty buffer probability for $\Lambda = 10^{-2}, 10^{-3}$, $P_c = 0.2\ W$, $\zeta = 1.2$, $\lambda = 1$, $n = 500$, and $\epsilon_t = 10^{-3}$.

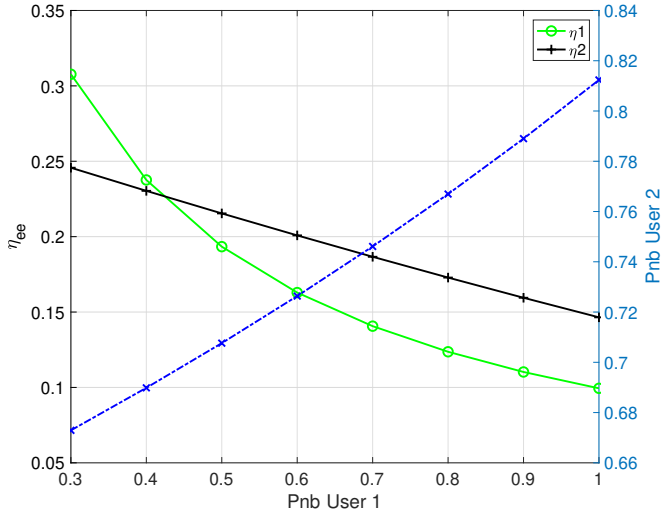

Fig. 7. Multi-user performance evaluation with EBP for $\epsilon_{u1} = 10^{-4}$, $\epsilon_{u2} = 0.1$, $\rho_1 = 6$ dB, $\rho_2 = 0$ dB, $\theta = 0.01$, $P_c = 0.2$, $\zeta = 1.2$, $\delta = 500$, $\lambda_2 = 0.5$ and $n = 500$.

the arrival rate of User 1 to a lower level reduces its NBP probability and improves the EEE of both users. Meanwhile, the NBP probability of User 2 increases when the buffer of User 1 is more busy. This happens because User 2 suffers from excess interference from User 1, which forces User 2 into reducing its transmission rate. Hence, packets accumulate in the buffer of User 2 which in turn becomes more congested.

In Fig. 8, we plot the optimum power allocation for maximizing the EEE as a function of the maximum delay δ in case of EBP and always full buffer where $\rho_{max} = 10$ dB. The target error outage probability is fixed at $\epsilon = 10^{-4}$. The plot shows that the optimal power allocation is significantly higher when the delay outage probability Λ is lower and when EBP is considered. The figure also depicts that Shannon's model does not render an accurate power allocation to maximize the EEE; in fact, it underestimates the optimum power allocation when compared to the finite blocklength model. The power gap

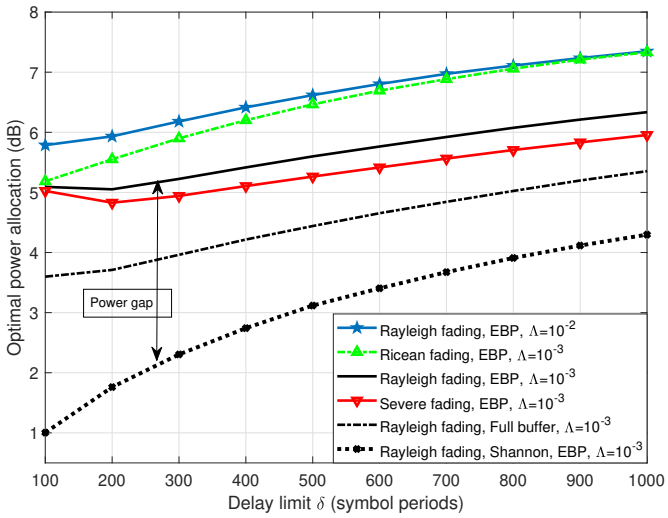

Fig. 8. Optimal power allocation vs δ with and without empty-buffer probability for $\Lambda = 10^{-2}, 10^{-3}$, $P_c = 0.2$, $\zeta = 1.2$, $\lambda = 1$, $\rho_{max} = 13$ dB and $\epsilon = 10^{-4}$.

ranges from 2 to 4 dB as shown in the figure. Thus, we can exploit the extra power allocation that results from considering empty buffer and applying the finite blocklength model in order to efficiently boost the EC. It is also observed that the optimal power allocation increases when the delay tolerance becomes higher. The intuition behind this is that when the network tolerates higher delays, it allows for improving the throughput by allocating higher power without wasting the network resources. This improvement occurs in the same way when considering empty buffer probability and when increasing the line of sight (e.g. Ricean fading where $m > 1$).

In Fig. 9, we illustrate the EEE gain of the EBP retransmission scenario. The system parameters are $\rho = 6$ dB, $\epsilon = 10^{-9}, \lambda = 0.5$. The figure shows that our proposed EBP scheme with ARQ enhances the EEE when compared to the classical EBP and full buffer models. In fact, the EEE of EBP model with ARQ is more than double of the normal EBP case when the delay constraint is very strict and the delay exponent approaches high values at $\theta > 0.1$. Thus, the EEE gain is more relevant for delay stringent networks. In this case, the EEE is upper bounded by 1.07 (bpcu/watt) as obtained from Theorem 4. The plot also compares different error allocation strategies for the first and second transmission rounds. It is obvious that the equal error allocation is not optimal enough to maximize the EEE. However, the minimum transmit power strategy highly approaches EEE optimality.

In Fig. 10, we depict the total power consumption accompanied by each scenario as function of the arrival rate. The network parameters are $\rho = 6$ dB, $n = 500, \epsilon = 10^{-9}, \theta = 0.01$, where λ is varied this time as shown on the figure axis. Although this effect is marginal for the full buffer model, the figure depicts that higher arrival rates consume more power as there are more packets to transmit. However, the power consumption is lower for the EBP model. Despite retransmissions which consume high power, the EBP model with retransmissions is the most power saving scheme, since the packets are transmitted with higher error probabilities for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3052965, IEEE Internet of Things Journal

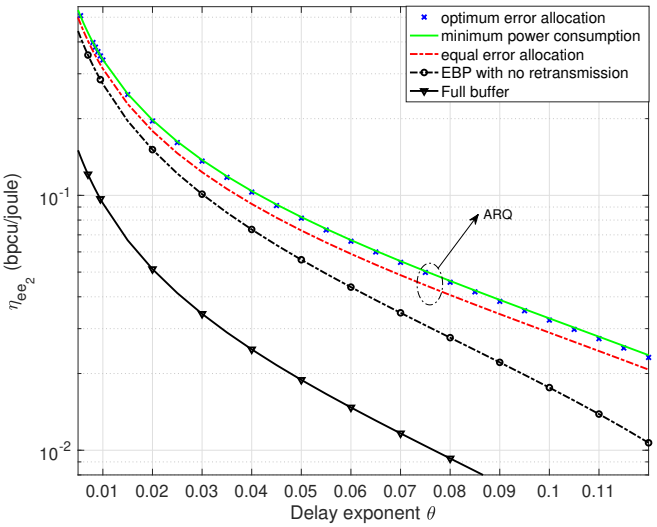

Fig. 9. Effective energy efficiency η_{ee} for different delay exponents θ, where $\rho = 6$ dB, $n = 500$, $\epsilon = 10^{-9}$, $P_c = 0.2$, $\zeta = 1.2$, $\lambda = 0.5$.

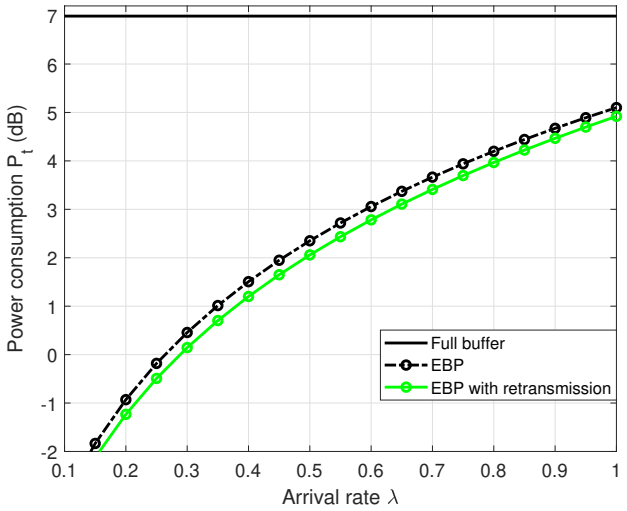

Fig. 10. Power consumption P_t for different arrival rates λ, where $\rho = 6$ dB, $n = 500$, $\epsilon = 10^{-9}$, $P_c = 0.2$, $\zeta = 1.2$, $\theta = 0.01$.

each single transmission which boosts the service rate and reduces the traffic congestion at the buffer and hence, the power consumption of the whole network.

Finally, Fig. 11 illustrates the normalized delay τ_n as a function of arrival rate λ for the EBP-ARQ scheme with two transmissions. The figure shows that the normalized delay τ_n diminishes as the network traffic becomes higher. This is because, when the queue becomes more congested, there are fewer chances for the buffer to become empty and hence, the opportunity for a second transmission disappears. Hence, the delay becomes only one transmission delay which is lower than the delay in case of two transmissions. It is noted that the delay becomes worse for lower reliability requirement as the error is also relaxed in the first transmission which leads to higher probability of occurrence for the second transmission and longer delay.

Moreover, for the same reliability constraint, boosting the

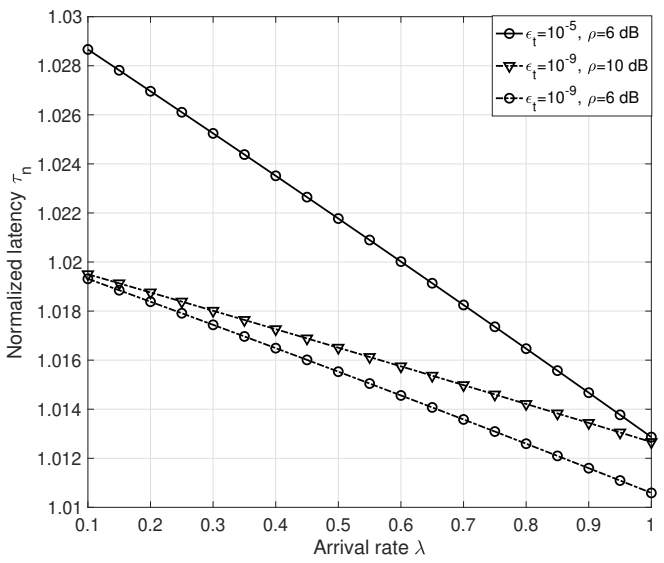

Fig. 11. Normalized delay τ_n for different arrival rates λ, where $n = 500$, $P_c = 0.2$, $\zeta = 1.2$, $\theta = 0.01$.

transmit power does not reduce latency. This is because for this high power, if the network becomes more congested, it slightly affects the non-EBP which maintains its low value and allows for second transmission which causes longer delay. The delay in its worst case is still only 3% higher than the delay of one transmission scheme. Hence, we obtain significantly higher EEE with only limited increase in the delay, while maintaining reliability at the same level.

VII. Conclusions

In this work, we presented a detailed analysis of the EEE for ultra reliable delay constrained networks in the finite blocklength regime. For Nakagami-m quasi-static fading channels, we proposed an approximation for the EC. Then we characterized the EEE maximizers in terms of optimum error probability and power allocation for the Rayleigh fading case. The results revealed that Shannon's model overestimates the EEE and underestimates the optimum power allocation when compared to the finite blocklength model. Further results indicated that allowing for larger delays significantly boosts EEE. We showed that the advantage of considering non-empty buffer probability and flexible transmission power is twofold since it significantly improves the EEE of networks operating in the finite blocklength regime and allows for retransmission of faulty packets with a significant boost in the EEE, reliability, and limited delay extension. For the EBP retransmission scenario, we derived the upper bound of the EEE and provided a low complexity solution for the optimization problem of maximizing the EEE and minimizing the power consumption. The solution showed that EBP model with retransmission is an ultra-reliable power saving scheme which improves the energy efficiency with limited increase in latency. Better performance and higher EEE gain could be achieved by applying Chase Combining (CC-HARQ) and Incremental redundancy (IR-HARQ) protocols [39]. This is

left as possible extension for this work along with the analysis of EEE in multi-user networks as in Fig. 7.

Appendix A
Proof of Lemma 1

Using (3) in (5), we attain

$$C_e(\rho,\theta,\epsilon) = -\frac{1}{n\theta}\ln\left(\frac{m^m}{\Gamma(m)}\int_0^\infty \left(\epsilon+(1-\epsilon)e^{-\theta nr}\right) z^{m-1}e^{-mz}dz\right). \tag{34}$$

From (2), we have

$$e^{-\theta nr} = e^{-\theta n\log_2(1+\rho z)} e^{\theta\sqrt{n(1-\frac{1}{(1+\rho z)^2})}Q^{-1}(\epsilon)\log_2 e}, \tag{35}$$

where

$$e^{-\theta n\log_2(1+\rho z)} = (1+\rho z)^{\alpha} \tag{36}$$

$$e^{\theta\sqrt{n(1-\frac{1}{(1+\rho z)^2})}Q^{-1}(\epsilon)\log_2 e} = e^{\beta\gamma}. \tag{37}$$

We resort to the Taylor expansion to obtain $e^{cx} = \sum_{n=0}^{\infty}\frac{(cx)^n}{n!}$. It follows from (35),(36) and (37) that the expression in (34) can be written as

$$EC(\rho_i,\theta,\epsilon) = -\frac{1}{n\theta}\ln\left[\int_0^\infty \epsilon\frac{m^m}{\Gamma(m)}\, z^{m-1}e^{-mz}dz + (1-\epsilon)\int_0^\infty \frac{m^m}{\Gamma(m)}(1+\rho_i z)^{\alpha}\sum_{n=0}^{\infty}\frac{(\beta\gamma)^n}{n!}z^{m-1}e^{-mz}dz\right]. \tag{38}$$

The infinite series in (38) can be truncated to a finite sum of terms and we evaluate the accuracy of the expression noting that the accuracy increases with the number of terms. But, it is noticed that when testing for different system parameters (N, ρ, θ, n), the accuracy for expanding 1 term is 92.7%, 2 terms is 99% and 99.9% for 3 terms only. Henceforth, in our analysis, 3 terms will be enough and (38) reduces to (9).

Appendix B
Proof of Theorem 1

The coefficients of Rayleigh channel are distributed according to the probability density function (PDF) $f_Z(z) = e^{-z}$. This corresponds to a Nakagami-m fading parameter $m = 1$. Applying the second order Taylor expansion to obtain $e^{\beta\gamma} = 1+(\beta\gamma)+\frac{(\beta\gamma)^2}{2}$, it follows from Theorem 1 that for Rayleigh distributed channels

$$\psi(\rho,\theta,\epsilon) = \epsilon+(1-\epsilon)\left[\int_0^\infty (1+\rho z)^{\alpha}e^{-z}\mathrm{d}z + \beta\int_0^\infty (1+\rho z)^{\alpha}\gamma e^{-z}\mathrm{d}z + \frac{\beta^2}{2}\int_0^\infty (1+\rho z)^{\alpha}\gamma^2 e^{-z}\mathrm{d}z\right]. \tag{39}$$

The first integral can be written as $e^{\frac{1}{\rho}}\rho^{\alpha}\Gamma\left(\alpha+1,\frac{1}{\rho}\right)$ By applying Laurent's expansion for γ [40], we obtain $\gamma \approx 1-\frac{1}{2(1+\rho z)^2}$. Hence, the second and third integrals can be written as $e^{\frac{1}{\rho}}\beta\rho^{\alpha}\left(\Gamma\left(\alpha+1,\frac{1}{\rho}\right)-\frac{\Gamma\left(\alpha-1,\frac{1}{\rho}\right)}{\rho^2}\right)$, and $e^{\frac{1}{\rho}}\frac{\beta^2}{2}\rho^{\alpha}\left(\Gamma\left(\alpha+1,\frac{1}{\rho}\right)-\frac{\Gamma\left(\alpha-1,\frac{1}{\rho}\right)}{\rho^2}\right)$, respectively leading to (11).

Appendix C
Proof of Theorem 2

Since the logarithmic term is dominant in the rate equation given in (2), it is quite straightforward to verify that the rate function has a negative second derivative for practical rate and SNR regions and therefore is concave in transmit power. The mathematical proof proceeds as follows. First, let $\phi = \frac{Q^{-1}(\epsilon)\log_2(e)}{\sqrt{n}}$ and note that ϕ should be a strictly positive parameter. Moreover, ϕ is less than unity for practical values of n and ϵ, where $n \geqslant 1$ and $\epsilon \leqslant 0.1$, which in fact are guaranteed for URLLC operation where n>100 and $\epsilon < 10^{-4}$ [1]. This dictates that the denominator of ϕ is higher than its numerator since $\sqrt{n}$ will be large enough to exceed $Q^{-1}(\epsilon)\log_2(e)$. Then from (2), we have

$$\frac{\partial r}{\partial\rho} = \frac{z}{(1+\rho z)\log 2} - \frac{\phi z}{(1+\rho z)^3\sqrt{1-\frac{1}{(1+\rho z)^2}}}, \tag{40}$$

$$\frac{\partial^2 r}{\partial\rho^2} = \frac{3\phi z^2}{(1+\rho z)^4\sqrt{1-\frac{1}{(1+\rho z)^2}}} + \frac{\phi z^2}{(1+\rho z)^6\left(1-\frac{1}{(1+\rho z)^2}\right)^{\frac{3}{2}}} - \frac{z^2}{(1+\rho z)^2\log 2}, \tag{41}$$

which is dominated by the negative term, since the other terms are multiplied by $\phi \lessapprox 1$ and raised to a high power in the denominator, and thus vanish faster. This firmly holds for non-extremely low SNR (i.e, $\geqslant -10$ dB) regions. Following a similar procedure as in [16] based on [41], we can conclude that the EEE in the finite blocklength regime is also a quasi-concave function of power and strictly concave in its upper contour.

Appendix D
Proof of Lemma 2

For $\rho = 0$, the achievable rate $r = 0$ and the numerator of (8) becomes 0. Applying L'Hopital's rule for the denominator, we have

$$\lim_{\rho\to 0}\frac{\rho}{\mathbb{E}[r]} = \lim_{\rho\to 0}\frac{1}{\mathbb{E}\left[z\left(\frac{1}{(1+\rho z)\ln 2}-\frac{Q^{-1}(\epsilon)\log_2(e)}{\sqrt{n}(1+\rho z)^3\gamma}\right)\right]} = 0. \tag{42}$$

Thus the denominator of (8) equals to P_c yielding 0 for the EEE.

For the second condition, the numerator of (8) is upper bounded by $-\frac{\ln\epsilon}{n\theta}$, while L'Hopital's rule for the denominator, we obtain

$$\lim_{\rho\to\infty}\frac{1}{\mathbb{E}\left[z\left(\frac{1}{(1+\rho z)\ln 2}-\frac{Q^{-1}(\epsilon)\log_2(e)}{\sqrt{n}(1+\rho z)^3\gamma}\right)\right]} = \infty. \tag{43}$$

Thus, the denominator of (8) tends to infinity which nulls the EEE. Hence, (19) holds as well under finite blocklength regime, which concludes the proof.

APPENDIX E
PROOF OF THEOREM 4

Since the denominator of the EEE does not depend on θ, so the problem is to define the limits of the EC. First, we define the limit of the EC for one transmission scenario and extremely strict delay constraint, where $\theta \to \infty$ as

$$\lim_{\theta\to\infty} C_{e1}(\epsilon) = \lim_{\theta\to\infty} \frac{-1}{n\theta} \ln\left(\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]\right) = 0, \tag{44}$$

which means that the EC vanishes as the delay constraint becomes infinitely strict and consequently, the EEE vanishes too. A similar procedure shows the same zero lower bound for the retransmission scenario. Next, we define the limit of the EC for loose delay constraint, where $\theta \to 0$ as follows

$$\lim_{\theta\to 0} C_{e1}(\epsilon) = \lim_{\theta\to 0} -\frac{f(\theta)}{g(\theta)} = \lim_{\theta\to 0} -\frac{\frac{\partial f(\theta)}{\partial\theta}}{\frac{\partial g(\theta)}{\partial\theta}} = \lim_{\theta\to 0} -\frac{\frac{\partial f(\theta)}{\partial\theta}}{n}, \tag{45}$$

which follows from L'Hopital rule, where $f(\theta) = \ln\left(\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]\right)$ and $g(\theta) = n\theta$. Differentiating $f(\theta)$ with respect to θ, we get

$$\begin{aligned}\frac{\partial f(\theta)}{\partial\theta} &= \frac{\frac{\partial}{\partial\theta}\left\{\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]\right\}}{\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]} \\ &= \frac{\frac{\partial}{\partial\theta}\int_0^\infty \left(\epsilon + (1-\epsilon)e^{-n\theta r}\right)e^{-z}dz}{\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]}.\end{aligned} \tag{46}$$

Applying Leibniz's rule, we obtain

$$\begin{aligned}\frac{\partial f(\theta)}{\partial\theta} &= \frac{\int_0^\infty \frac{\partial}{\partial\theta}\left(\epsilon + (1-\epsilon)e^{-n\theta r}\right)e^{-z}dz}{\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]} \\ &= \frac{-n(1-\epsilon)\,\mathbb{E}_Z\left[re^{-n\theta r(\epsilon)}\right]}{\mathbb{E}_Z\left[\epsilon + (1-\epsilon)e^{-n\theta r(\epsilon)}\right]}.\end{aligned} \tag{47}$$

Plugging back into (45), we reach

$$\begin{aligned}\lim_{\theta\to 0} C_e(\epsilon) &= \lim_{\theta\to 0} -\frac{\frac{-n(1-\epsilon)\,\mathbb{E}_Z\left[re^{-n\theta r(\epsilon)}\right]}{\mathbb{E}_Z\left[\epsilon+(1-\epsilon)e^{-n\theta r(\epsilon)}\right]}}{n} \\ &= (1-\epsilon)\,\mathbb{E}_Z\left[r(\epsilon)\right] = (1-\epsilon)r_0,\end{aligned} \tag{48}$$

which represents the upper bound throughput of the finite blocklength transmission when no delay constraint is imposed. Following the same procedure, we can deduce that the EC of the EBP ARQ scenario is given by the numerator of Theorem 4.

APPENDIX F
PROOF OF THEOREM 5

The proof for the pseudo-convexity of $p_{nb}^{'}$ in ϵ_1 goes as follows. At the first glance, it is possible to prove that κ is a concave function in ϵ_1. First, let

$$r(\epsilon) \approx \log_2(1+\rho|h|^2) - \mu Q^{-1}(\epsilon), \tag{49}$$

where $\mu = \frac{\log_2(e)}{\sqrt{n}}\sqrt{1 - \frac{1}{\left(1+\rho|h|^2\right)^2}}$. Applying the derivatives of the inverse Q-function from [6], we obtain the derivatives of the rate expectations with respect to ϵ_1 as

$$\frac{\partial r_1}{\partial\epsilon_1} = \mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}, \tag{50}$$

$$\frac{\partial^2 r_1}{\partial\epsilon_1^2} = -\mathbb{E}\left[\mu\right]2\pi Q^{-1}(\epsilon_1)e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}, \tag{51}$$

$$\frac{\partial r_2}{\partial\epsilon_1} = -\frac{\epsilon}{\epsilon_1^2}\mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}, \tag{52}$$

$$\begin{aligned}\frac{\partial^2 r_2}{\partial\epsilon_1^2} &= \frac{\epsilon}{\epsilon_1^2}\mathbb{E}\left[\mu\right]2\pi Q^{-1}(\epsilon_1)e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}} \\ &\quad + \mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}\frac{2\epsilon}{\epsilon_1^3} \\ &= -\frac{\epsilon}{\epsilon_1^2}\frac{\partial^2 r_1}{\partial\epsilon_1^2} + \mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}\frac{2\epsilon}{\epsilon_1^3}.\end{aligned} \tag{53}$$

Then, we obtain the second derivative of κ w.r.t ϵ_1 as follows

$$\frac{\partial\kappa}{\partial\epsilon_1} = -r_1 + \frac{\partial r_1}{\partial\epsilon_1}(1-\epsilon_1) + \frac{1}{2}\left(\epsilon_1\frac{\partial r_2}{\partial\epsilon_1} + r_2\right) \tag{54}$$

$$\begin{aligned}\frac{\partial^2\kappa}{\partial\epsilon_1^2} &= -2\frac{\partial r_1}{\partial\epsilon_1} + \frac{\partial^2 r_1}{\partial\epsilon_1^2}(1-\epsilon_1) + \frac{1}{2}\left(2\frac{\partial r_2}{\partial\epsilon_1} + \epsilon_1\frac{\partial^2 r_2}{\partial\epsilon_1^2}\right) = \\ &-2\frac{\partial r_1}{\partial\epsilon_1} + \frac{\partial^2 r_1}{\partial\epsilon_1^2}(1-\epsilon_1) + \frac{1}{2}\left\{-2\frac{\epsilon}{\epsilon_1^2}\mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}\right. \\ &\left. -\frac{\epsilon}{\epsilon_1}\frac{\partial^2 r_1}{\partial\epsilon_1^2} + \epsilon_1\mathbb{E}\left[\mu\right]\sqrt{2\pi}e^{\frac{\left(Q^{-1}(\epsilon_1)\right)^2}{2}}\frac{2\epsilon}{\epsilon_1^3}\right\} \\ &= -2\frac{\partial r_1}{\partial\epsilon_1} + \frac{\partial^2 r_1}{\partial\epsilon_1^2}(1-\epsilon_1-\frac{\epsilon}{2\epsilon_1}).\end{aligned} \tag{55}$$

Note that practically, the term $\frac{\epsilon}{2\epsilon_1}$ is close to zero since $\epsilon_1 >> \epsilon$, and hence, $(1-\epsilon_1) > \frac{\epsilon}{2\epsilon_1}$. Since the derivatives in (50) and (51) are strictly negative, we can deduce that the second derivative of κ in (55) renders a strictly negative value. Therefore, κ is strictly concave in ϵ_1. Back to equation (24), the denominator of the non-empty buffer probability is a linear decreasing function of κ, which indicates that the denominator is a convex function in ϵ_1. The numerator can be proven to be a convex decreasing function of T which means that the numerator is a convex function in ϵ_1. First we obtain the derivatives of the numerator of $p_{nb}^{'}$ with respect to T as follows

$$\frac{\partial\ num\left\{p_{nb}^{'}\right\}}{\partial\kappa} = \frac{2\kappa - 4\lambda}{2\sqrt{\kappa^2 + 4\lambda(r_o - \kappa)}} - 1, \tag{56}$$

$$\frac{\partial^2\ num\left\{p_{nb}^{'}\right\}}{\partial\kappa^2} = \frac{4\lambda(r_o - \lambda)}{\left(\kappa^2 + 4\lambda(r_o - \kappa)\right)^{\frac{3}{2}}}, \tag{57}$$

which gives a strictly positive value and hence, the numerator of of $p_{nb}^{'}$ is convex in κ. Since the numerator of $p_{nb}^{'}$ is a non-increasing convex function in κ and κ is a concave function of ϵ_1, it follows from [36] that it is a convex function in ϵ_1, while the numerator is also a negative function. Hence and according to Proposition 2.9 in [28], minimizing $p_{nb}^{'}$ is a psuedo-convex fractional program.

REFERENCES

[1] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[2] C. Liu, Y. Shen, and C. Lee, "Energy-Efficient Activation and Uplink Transmission for Cellular IoT," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 906–921, Feb 2020.

[3] C. Kurnaz and B. Engiz, "Monitoring and Assessment of Electromagnetic Pollution in Samsun (Turkey)," in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, June 2016, pp. 219–222.

[4] C. H. Liu and K. L. Fong, "Fundamentals of the Downlink Green Coverage and Energy Efficiency in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3271–3287, Dec 2016.

[5] R. Devassy, G. Durisi, P. Popovski, and E. G. Strom, "Finite-blocklength Analysis of the ARQ-protocol Throughput over the Gaussian Collision Channel," *ISCCSP 2014*, pp. 173–177.

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

[7] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing Resource Allocation in the Short Blocklength Regime for Ultra-Reliable and Low-Latency Communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, Jan 2019.

[8] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[9] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[10] 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, 2003.

[11] B. Makki, C. Fang, T. Svensson, M. Nasiri-Kenari, and M. Zorzi, "Delay-Sensitive Area Spectral Efficiency: A Performance Metric for Delay-Constrained Green Networks," *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2467–2480, 2017.

[12] 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 Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5827–5840, 2018.

[13] M. Shehab, E. Dosti, H. Alves, and M. Latva-aho, "On the Effective Capacity of MTC Networks in the Finite Blocklength Regime," in *EUCNC 2017*, Oulu, Finland, Jun. 2017.

[14] M. Sinaie, A. Zappone, E. A. Jorswieck, and P. Azmi, "A Novel Power Consumption Model for Effective Energy Efficiency in Wireless Networks," *IEEE Wireless Communications Letters*, vol. 5, no. 2, pp. 152–155, 2016.

[15] M. Shehab, E. Dosti, H. Alves, and M. Latva-aho, "On the Effective Energy Efficiency of Ultra-reliable Networks in the Finite Blocklength Regime," in *2017 International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2017, pp. 275–280.

[16] L. Musavian and Q. Ni, "Effective Capacity Maximization with Statistical Delay and Effective Energy Efficiency Requirements," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3824–3835, July 2015.

[17] C. She and C. Yang, "Energy Efficiency and Delay in Wireless Systems: Is Their Relation Always a Tradeoff?" *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7215–7228, 2016.

[18] 3GPP, "3GPP TS 38.300: Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2;," *Rel 15*, vol. 15.5.0, pp. 1–97, 2019.

[19] P. Larsson, J. Gross, H. Al-Zubaidy, L. K. Rasmussen, and M. Skoglund, "Effective Capacity of Retransmission Schemes: A Recurrence Relation Approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4817–4835, Nov 2016.

[20] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast HARQ Over Finite Blocklength Codes: A Technique for Low-Latency Reliable Communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 194–209, 2019.

[21] M. Shehab, H. Alves, and M. Latva-aho, "Effective Capacity and Power Allocation for Machine-Type Communication," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4098–4102, April 2019.

[22] N. H. Mahmood, O. A. Lopez, H. Alves, and M. Latva-aho, "A Predictive Interference Management Algorithm for URLLC in Beyond 5G Networks," *IEEE Communications Letters*, pp. 1–1, 2020.

[23] R. Jurdi, S. R. Khosravirad, and H. Viswanathan, "Variable-rate ultra-reliable and low-latency communication for industrial automation," in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, March 2018, pp. 1–6.

[24] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*. Copyright 2000 John Wiley & Sons, Inc., 2005.

[25] L. Musavian and T. Le-Ngoc, "QoS-based power allocation for cognitive radios with AMC and ARQ in Nakagami-m fading Channels," in *Trans. Emerging Tel. Tech.*, vol. 27, 2014, pp. 266–277.

[26] N. Petreska, H. Al-Zubaidy, R. Knorr, and J. Gross, "Bound-based power optimization for multi-hop heterogeneous wireless industrial networks under statistical delay constraints," *Comput. Networks*, vol. 148, pp. 262–279, 2019. [Online]. Available: https://doi.org/10.1016/j.comnet.2018.09.009

[27] A. Helmy, L. Musavian, and T. Le-Ngoc, "Energy-Efficient Power Adaptation over a Frequency-Selective Fading Channel with Delay and Power Constraints," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4529–4541, September 2013.

[28] A. Zappone and E. Jorswieck, *Energy efficiency in wireless networks via fractional programming theory*. Foundations and Trends in Communications and Information Theory,, 2015, vol. 11.

[29] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-Efficient Resource Allocation in Multi-Cell OFDMA Systems with Limited Backhaul Capacityâs ," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3618–3631, 2012.

[30] S. I. Abramowitz M, *Handbook of mathematical functions*. New York: Dover, 1965.

[31] I. S. Gradshteyn and I. M. Ryshik, *Table of Integrals, Series, and Products*, 5th ed. London: Academic Press, 1994.

[32] E. Dosti, U. L. Wijewardhana, H. Alves, and M. Latva-aho, "Ultra Reliable Communication via Optimum Power Allocation for Type-I ARQ in Finite Block-Length," in *IEEE ICC 2017*, Paris, France, May 2017, pp. 5168–5173.

[33] C. F. Lai, Y. C. Chang, H. C. Chao, M. S. Hossain, and A. Ghoneim, "A Buffer-Aware QoS Streaming Approach for SDN-Enabled 5G Vehicular Networks," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 68–73, 2017.

[34] Z. Liu, B. Liu, and C. W. Chen, "Buffer-Aware Resource Allocation Scheme with Energy Efficiency and QoS Effectiveness in Wireless Body Area Networks," *IEEE Access*, vol. 5, pp. 20 763–20 776, 2017.

[35] Z. Liu, B. Liu, and C. Chen, "Transmission-Rate-Adaption Assisted Energy-Efficient Resource Allocation With QoS Support in WBANs," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5767–5780, Sept 2017.

[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[37] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer Publishing Company, Incorporated, 2014.

[38] L. Kundu, G. Xiong, and J. Cho, "Physical Uplink Control Channel Design for 5G New Radio," in *2018 IEEE 5G World Forum (5GWF)*, July 2018, pp. 233–238.

[39] E. Dosti, M. Shehab, H. Alves, and M. Latva-Aho, "Ultra reliable communication via optimum power allocation for harq retransmission schemes," *IEEE Access*, vol. 8, pp. 89 768–89 781, 2020.

[40] K. I. R. Rodriguez and J. Gilman, *Complex Analysis: In the Spirit of Lipman Bers.* Springer, 2012.

[41] L. Liu, Y. Yi, J. F. Chamberland, and J. Zhang, "Energy-Efficient Power Allocation for Delay-Sensitive Multimedia Traffic Over Wireless Systems," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2038–2047, Jun 2014.