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.2022.3157956, IEEE Internet of Things Journal

Effective Energy Efficiency and Statistical QoS Provisioning under Markovian Arrivals and Finite Blocklength Regime

Fahad Qasmi, Mohammad Shehab, Hirley Alves, and Matti Latva-aho

***Abstract*—In this paper, we evaluate the Effective Energy Efficiency (EEE) and propose delay-outage aware resource allocation strategies for energy-limited IoT (Internet of Things) devices under the finite blocklength (FBL) regime. The EEE is a cross-layer model, measured by the ratio of Effective Capacity to the total consumed power. To maximize the EEE, there is a need to optimize transmission parameters such as transmission power and rate efficiently. Whereas it is quite complex to study the impact of transmission power, or rate alone, the complexity is aggravated by the simultaneous consideration of both variables. Hence, we formulate power allocation (PA) and rate allocation (RA) optimization problems individually and jointly to maximize EEE. Furthermore, we investigate the performance of the EEE under constant and random arrivals, where statistical QoS constraints are imposed on buffer overflow probability. Using effective bandwidth and effective capacity theories, we determine the arrival rate and the required service rate that satisfy the QoS constraints. After that, we compare the performance of different iterative algorithms such as Dinkelbach's and Cross Entropy, which guarantee the convergence for the optimal solution. By numerical analysis, the influence of source characteristics, fixed transmission rate, error probability, coding blocklength, and QoS constraints on the throughput are identified. Our analysis reveals that the joint PA and RA is the optimal resources allocation strategy for maximizing the EEE in the presence of constant and random data arrivals. Finally, the results illustrate that the modified Dinkelbach's algorithm has high performance and low complexity compared to others.**

***Index Terms*—Finite Blocklength, Effective Capacity, Traffic, Radio Resource Management, IoT, MTC, EEE, QoS.**

I. Introduction

The massive machine type communication (mMTC) and URLLC (Ultra-Reliable Low-Latency Communication) are the next-generation cellular network use cases that will unleash emerging applications and industry verticals. This will lead organisations to transform their business models and digitalise the industrial process towards IoT (Internet of Things). These two new communications modes are dedicated to IoT devices, operating without human intervention [1]. It is predicted that more than 50 billion devices will be connected via cellular access technologies by 2025 [2].

The mMTC is a network service that supports a large number of connected power-limited IoT devices that send or receive messages, with some tolerable delay requirements. These IoT are usually used for sensing, and data transmission for smart monitoring, automation and infrastructure of buildings, smart agriculture and fleet management [3].

In this regard, the existing literature on IoT communication focuses on several challenges including resource allocation, power minimization, ensuring reliability and latency [4]. Current cellular network is mainly designed and optimized based on human-centric communication. It is relatively characterized by a smaller number of devices with high downlink data rates and large packet size. Meanwhile, in IoT ecosystem generally involves numerous devices with mostly low uplink rate and sporadic transmissions of small packets. Hence, existing cellular networks would be unable to handle this kind of traffic because of a large number of IoT overwhelmingly attempting to access spectrum in coordinated way, which leads to signal overhead and congestion in the network [5]. Besides, in order to achieve low latency and due to the nature of IoT sensors' data, it will rely on short transmission packets, composed of a smaller number of bytes. Therefore, they require fewer radio resources, which may lead to packet scheduling problems and reducing capacity of the network [6].

A. Literature Review

IoT devices' traffic is often characterized as (non) real-time, periodic, sporadic, event-driven, bursty, and homogeneous in nature [7], [8]. In many IoT-type applications, messages are short, demanding a novel framework to handle finite block communication. Fortunately, during the last decade, especially after the seminal work in [9], extensive progress has been made within the information theory community to address the issues of transmitting short packets and explore theoretical principles governing the short-packet transmission and possess metrics that allow to assess their performance [10]. The authors in [9] proposed an approximation for the achievable coding rate under finite blocklength regime and then investigated the performance gap that appears when applying finite blocklength rate equations. Additionally, in [11], the authors proposed a maximum communication rate as a function of blocklength and error outage probability, in contrast to Shannon's coding theorem that attains error-free communication when the blocklength goes to infinity [12]. In [13], authors investigated the tradeoff between the sum rate and the error probability in finite blocklength regime. The results proved that minimizing the per-user error probability plays a critical role in achieving high throughput under delay-constrained scenarios.

Authors are with the Centre for Wireless Communications (CWC), University of Oulu, Finland. {firstname.lastname}@oulu.fi.

This research has been financially supported by Academy of Finland, 6Genesis Flagship (Grant n.318937) and ee-IoT (Grant n.319008), and Academy Professor (Grant n.307492).

Moreover, the traffic is generated in the IoT ecosystem requires service guarantees in the time-varying wireless channel, which may cause severe QoS violations due to the random changes in the environment or obstacles [14]. Due to random variations in the radio channels, deterministic QoS constraints are usually difficult to guarantee for delay-sensitive services over wireless networks. Hence, in [15] authors proposed to shift the concept from deterministic to statistical QoS guarantees using the concept of effective capacity, which ensures latency guarantees with the maximum constant arrival of data served by the random wireless channel. Effective capacity has been extensively applied over the past few years to evaluate the trade-off among the reliability, latency, security, and energy efficiency evinced by a recent survey [16]. Meanwhile, the delay-Sensitive Area Spectral Efficiency metric in [17] only considered data transmission delay in the performance analysis under coverage area. However, it does not consider the statistical aspect that quantifies QoS in terms of delay outage probability and the maximum delay bound as in the effective capacity metric. In recent contributions such as [18] the authors evaluated the performance of the effective capacity together with Markovian arrival traffic, evaluating the end-to-end performance based on arrival and services processes. Besides, the authors assumed a fixed rate transmission over the Rayleigh fading channel modelled as discrete-time Markov chain. These channel models identify the level of reliability and latency that each optimum transmission rate and maximize the throughput that tolerates higher traffic arrival rate generated by sources. In [19], the authors incorporated effective capacity with finite blocklength to analyse the trade-off between reliability and latency.

In the literature, there are two energy efficiency models that attracted extensive attention for cellular networks. The first is based on network capacity per unit energy consumption, and the second is effective capacity per unit energy consumption model [20], [21]. The network capacity describes a theoretical performance of upper bound throughput, which does not consider delay QoS requirements in practical applications. Therefore, it is mainly used for evaluating the performance of delay insensitive services [22], [23]. However, the network capacity analysis formulation may not be accurate for supporting delay-sensitive IoT systems. The energy efficiency issue in mMTC was studied in [23], where the authors investigated an analytical framework for energy efficiency that can be defined as the ratio of achievable rate to energy consumption, that comprises radiator and static circuit powers. In [21], authors evaluated the effective energy efficiency (EEE) for delay-sensitive networks in the finite blocklength regime and propose an optimum power allocation strategy. Results confirm that Shannon's model underestimates the optimum power when compared to the exact finite blocklength model. Authors in [24] addressed a joint power and rate allocation problem for minimizing the time required for the concurrent transmission of IoT while satisfying delay, reliability, and energy consumption requirements. The authors of [25] showed that in wireless systems, the relation between EEE and delay is not always a trade-off. They concluded that there is an EEE-delay non-trade-off region where the service rate and power consumption are linearly related. In [26] the rate allocation problem was discussed for downlink cellular networks with Rayleigh fading and stringent reliability constraints. The allocated rate depends on the target reliability, average statistics of the signal, interference, and the number of antennas at the receiver. In all previous works on EEE, the buffer was assumed to be always full. Therefore, a transmitter is considered to be always ON and consumes more power during communication. Instead, a more realistic approach accounts for the probability of an empty buffer due to a constant and random data arrival traffic during a communication. In [27], the non-empty buffer probability model was known as EEE boost for transmission of large packets. This model gives a more realistic estimation of effective capacity due to assuming the probability of emptying the buffer during transmission with constant arrival rate. Recently, in [21] authors proved that this power consumption model is valid for finite blocklength regime. The authors [28] proposed a EEE model based on empty and non-empty buffer for large packets transmission. This model proved justifiable characteristics to apprehend random arrival traffics and always attains higher EEE than the full buffer scenario.

B. Contribution

Different from the above works, herein, we study the effective energy efficiency under finite blocklength to satisfy a delay-outage constraint when the arrival traffics are sporadic under fading channel. We consider a power-limited IoT system where CSI acquisition procedures are quite expensive[1]. Thus, the transmitter communicates at a fixed transmission rate. We propose a power consumption model which is motivated from non-empty buffer probability (NBP) [28]. Our proposed optimum power and rate allocation strategies are based on buffer condition and available resources of the channel without degrading the performance. Moreover, We evaluate the EEE of NBP model and compare it to the basic FBL model. The results show that NBP is a more accurate and realistic model to captures the sporadic behaviour of IoT devices traffics. The main contributions of the paper are as follows:

- We propose power, rate and joint power and rate allocation strategies that meet the stringent delay-outage constraints of the random traffic, while maximizing the EEE.
- We derive an upper and lower bound of the EEE to prove the validity of this consumption model for constant and sporadic traffic arrivals.
- We prove that the EEE is quasi-concave as a function of the signal to noise ratio (SNR).
  We resorted to a non-empty buffer probability-based energy consumption model under sporadic traffic arrivals with short packets transmission.
- We solve the optimization problems for optimum power and rate allocation at each transmission with a power

[1]Channel acquisition overhead is one of the key characteristics of modern wireless systems which critically affects the throughput, reliability, latency and energy efficiency constraints. Accurate channel estimation could be acquired with the increase in the training phase; however, increasing the number of pilot symbols in the training phase reduces the effective data rate and wastes resources that could be used for data transmission [29].

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.2022.3157956, IEEE Internet of Things Journal

constraint via different low complexity algorithms such as Dinkelbach's, Cross Entropy optimization and Matlab function $fmincon$. The solution proved that minimizing the NBP jointly reduces the power consumption and maximizes the EEE.

- Lastly, we propose the integration of the Golden Section Search method with conventional Dinkelbach's algorithm to improve the convergence speed and compare its convergence to the other algorithms until they attain the optimal solution.

C. Outline

The remainder of this paper is organized as follows. In Section II-A, we describe the channel model. Section II-B revisits key concepts on theories of effective bandwidth and effective capacity. Later, in Section II-C, we define effective rate under finite blocklength regime. In Section II-D we discuss the traffic model for IoT applications through discrete-time Markov source. Section III explains EEE for both constant and random arrivals. The formulation of the EEE maximization problems and discussion of different algorithms for obtaining solutions and computational complexity analysis of algorithms are depicted in Section IV. Section V illustrates the numerical results. Finally, concluding remarks are stated in Section VI.

TABLE I
IMPORTANT ABBREVIATIONS AND SYMBOLS.

Symbol	Definition
EEE	effective energy efficiency
NBP	non-empty buffer probability
FBP	full buffer probability
h	fading coefficient
D	queueing delay
d	maximum delay
P_{nb}	non-empty buffer probability
θ	delay exponent
a^*	effective bandwidth
R_E	effective capacity / effective rate
m	fading parameter
ϵ	outage probability
ρ	average signal-to-noise ratio
r	achievable channel rate
n	blocklength
λ	arrival rate
λ_{avg}	average arrival rate
p	source arrival probability
ξ	inverse drain efficiency
P_c	hardware power dissipated
DK	Dinkelbach's Algorithm
CE	Cross Entropy Algorithm
PA	power allocation
RA	rate allocation

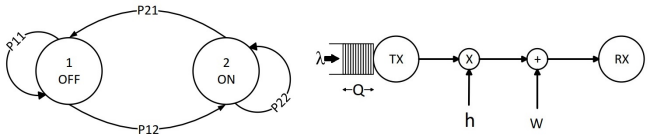


Fig. 1. System Model, where λ is the arrival rate at the source, and Q is the length of the queue at the transmitter. The channel coefficient is denoted as h and AWGN noise as **w.**

Notations: Throughout this paper, the expectation operator is denoted as $\mathbb{E}[\cdot]$. The Gamma function is denoted as $\Gamma(.)$ [30, Ch 6, 6.1.1], and $\Gamma(.,.)$ is the upper incomplete gamma function [30, Ch 6, 6.5].

II. System Model

A. Channel Model

We consider a point-to-point network in which a single energy-limited sensor transmits short packets to a common aggregator using a wireless communication system. It is assumed that the generated data is initially stored in the buffer before transmission over a wireless channel, as shown in Fig. 1. The channel input-output relation in a block can be expressed as $\mathbf{y} = h\mathbf{x} + \mathbf{w}$, where $\mathbf{x}, \mathbf{y}, \in \mathbb{C}^{\mathbf{n}}$ are the complex-valued vectors of channel input and output, and $h \in \mathbb{C}$ denotes quasi-static Nakagami-m block fading channel coefficient. Thus, the channel gain remains constant over the blocklength n and changes independently between block-to-block, which is assumed to be independent and identically distributed (i.i.d). The n is considered equal to the channel's coherence time due to low mobility or not transmitting often. Herein, $\mathbf{w}$ represents the additive Gaussian noise vector whose values are independently and identically distributed (i.i.d), complex, circularly symmetric random variables with unit mean and variance σ^2. Finally, we assume that channel state information (CSI) is only available at the receiver. Therefore, the receiver reliably detects the signal with limited errors. Additionally, as in [21], we aim to provide a performance benchmark for the EEE[2] of these networks, where the cost evaluation of CSI at the receiver is beyond the scope of our work.

B. Statistical Delay Constrained Analysis

The data generated by random sources is stored in a First In First Out (FIFO) buffer at the transmitter before transmission. Thus, delay may occur in the transmission because of the long waiting time of the data in the buffer. Therefore, it is assumed that the transmitter operates under statistical QoS constraints that would be applied with the purpose of restricting buffer overflow probability, which is defined as [31]

[2]We mainly concentrate our discussion on the EEE framework for energy-limited IoT. Our main objective is to evaluate the impact of sporadic traffic for different consumption models. For cost evaluation of CSI at the receiver (CSIR), power consumption models might become more complex. Thus, it might masquerade some of the effects on EEE. Therefore, we assume the ideal case where perfect CSIR is available at the receiver. Our results obtained under this assumption provide upper bounds of the performance and such an assumption allows us to focus on the impact of sporadic traffics in different consumption models used in EEE framework. However, the cost evaluation of CSIR under finite blocklength regime is interesting and is a possible extension of our work along with considering random access, channel estimation, and user detection in massive MTC scenarios.

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.2022.3157956, IEEE Internet of Things Journal

$$\lim_{q\to\infty} \frac{\ln \Pr\{Q \geq q\}}{q} = -\theta, \tag{1}$$

where the length of the stationary queue is represented by Q in steady state, θ denotes the decay rate of the tail distribution of the queue length, and q is the buffer threshold. For a relatively large q, the probability of buffer overflow can be approximated as

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

where $P_{nb} = \Pr\{Q > 0\}$ is the probability of non-empty buffer.[3] The statistical QoS guarantees are characterized by the QoS exponent θ. It is noted that θ controls the exponential decay rate of probability of buffer overflow; thus larger values of θ indicate a strict bound on the delay violation probability, which corresponds to more stringent QoS constraint and the system cannot tolerate delay. On the other hand, lower values of θ imply loose QoS guarantees imposed by the transmitter and therefore, the system can tolerate larger delay. In the given setting, the delay violation probability is also characterized to decay exponentially and is approximated by

$$\Pr\{D \geq d\} \approx P_{nb}\, e^{-\theta a^*(\theta) d}, \tag{3}$$

where, d is the delay threshold, D represents queueing delay at steady state and $a^*(\theta)$ denotes effective bandwidth [31].

Effective bandwidth characterizes the minimum constant service rates required to support the random arrival of data in the buffer constrained to some statistical QoS requirements, namely buffer violation probability. Let the time-accumulated arrival process at instant t be $A(t) = \sum_{k=1}^{t} a(k)$. Then the effective bandwidth is defined as [31]

$$a^*(\theta) = \lim_{t\to\infty} \frac{1}{\theta t} \log_e \mathbb{E}\{e^{\theta A(t)}\}. \tag{4}$$

Effective capacity (R_E) is the dual concept of effective bandwidth. It defines the maximum constant arrival rate, which is supported by given time-varying service process in order to guarantee a statistical QoS requirement specified by the QoS exponent θ. The effective capacity for a given QoS exponent is obtained from [32]

$$R_E(\theta) = -\lim_{t\to\infty} \frac{1}{\theta t} \log_e \mathbb{E}\{e^{-\theta S[t]}\}, \tag{5}$$

where $S[t] \triangleq \sum_{k=1}^{t} R_i$ is the time-accumulated service process and $\{R_i, i = 1, 2, \cdots\}$ is the discrete-time stationary and ergodic stochastic service process. It is noted that in the remainder of the paper, R_E will mention as the effective rate rather than the effective capacity when the service rates are equal to the approximate achievable coding rates in the finite blocklength regime.

[3]The probability of non-empty buffer if often approximated by the ratio of the average arrival rate λ_{avg} to the average service rate [23].

C. Effective Rate under Finite Blocklength Regime

In the information-theoretic analysis for infinite blocklength communication, k bits of information are encoded into the n symbol codeword, which is conveyed to the decoder via a wireless noisy channel, with coding rate $r = k/n$. When the codeword length increases without bound, we can achieve reliable transmission with no decoding errors for any transmission rate less than the Shannon's channel capacity. However, in IoT deployments, packets are short due to latency requirements or by system design and application requirements. In that sense, [9] attained an accurate approximation that characterizes the error probability in finite blocklength regime. The short packets are exchanged between nodes with new achievable coding rate r in the presence of target error probability ϵ and n; error probability is minimal but not vanishing. The normalised r in bits per channel uses (bpcu) is expressed as

$$r(\rho) \approx \log_2(1+\rho|h|^2) - \frac{Q^{-1}(\epsilon)\log_2(\epsilon)}{\sqrt{n}} \sqrt{1 - \frac{1}{(1+\rho|h|^2)^2}}, \tag{6}$$

where $Q(.) = \int_{.}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt$ is the Gaussian Q-function, Q^{-1} denotes its inverse, ρ represents the average signal to noise ratio (SNR). Notice that the noise is normalized so that the ρ denotes the transmit power and $|h|^2$ is the squared envelope of the channel fading coefficients. The fading coefficients are expressed by the random variable $Z = |h|^2$, which is gamma distributed with a probability density function (PDF) given as [23] $f_Z(z) = \frac{m^m z^{m-1}}{\Gamma(m)} e^{-mz}$, low values of m represent severe fading, whereas high values of m show the presence of line of sight (LoS) and $m = 1$ characterizes Rayleigh fading. From (6) we known that there is a performance gap between Shannon's capacity and finite blocklength that diminishes by increasing n [9]. It is practically assumed that the transmitter sends information at a fixed rate because it does not know the channel condition, or even know the channel. So due to many complexities in varying transmission rate for each fading block, the transmitter sends information at a fixed rate. Hence, based on [33], quasi-static fading channels the outage probability is

$$\mathbb{E}_{|h|^2}[\epsilon] = \int_0^{\infty} Q\left(\frac{\log_2(1+\rho|h|^2) - r}{\sqrt{\frac{1}{n}\left(1 - \frac{1}{(1+\rho|h|^2)^2}\right)} \log_2 e} \right) f_H(h) dh. \tag{7}$$

The outage probability in (7) does not have a closed form solution, but it can tightly be approximated using linearization of the Q-function as in (8) (on the top of the next page), where $\alpha = \frac{2^r - 1}{\rho}$, $\vartheta = \alpha + \sqrt{\frac{\pi}{2}\kappa^{-2}}$, $\varrho = \alpha - \sqrt{\frac{\pi}{2}\kappa^{-2}}$ and $\kappa = \sqrt{\frac{n\rho^2}{2\pi}}(e^{2r} - 1)^{-\frac{1}{2}}$ [33].

Remark 1. *Note that CSI at the transmitter (CSIT) may not be affordable for massive MTC scenario due to the fact that sensor is mostly uplink oriented and uses sporadic transmissions (e.g., due to event-driven traffic) with an often small transmission rate. Due to these characteristics the devices do not undergo scheduling procedures (e.g. grant free random access) [34]. Moreover, such devices have limited power and*

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.2022.3157956, IEEE Internet of Things Journal

$$\epsilon = \frac{\kappa}{\sqrt{2\pi}}\Big\{\alpha\Big(\Gamma(m, m\varrho) - \Gamma(m, m\vartheta)\Big) + (\Gamma(1+m, m\vartheta) - \Gamma(1+m, m\varrho))\Big\} + \frac{1}{2}\Big\{\Gamma\left(m, m\vartheta\right) + \Gamma\left(m, m\varrho\right)\Big\}. \tag{8}$$

baseband signal processing capabilities. As a result acquiring CSIT would drain energy resources significantly [35]. Secondly, the receiver can acquire CSI not only from the dedicated training sequence, but also from the codeword itself (combined estimation and decoding). Lastly, a small estimation error at the transmitter can cause severe interference and outages in transmission, whereas a small estimation error at the receiver simply adds a small additional noise term in the decoding process [29].

As in [36], we assume that the instantaneous transmission rate is fixed in each fading block channel. Hence, the ϵ varies with each block fading realization. In the case of transmission outage, date rate will be zero otherwise nr. Under these assumptions, instantaneous service rate in each block is

$$R_i = \begin{cases} 0 & \text{with prob. } \epsilon \\ nr & \text{with prob. } 1-\epsilon. \end{cases} \tag{9}$$

Now we are able to express the effective rate (in bits per channel use), under fixed rate transmissions and finite blocklength, as [36]

$$R_E(\theta, n, r, \rho) = -\frac{1}{n\theta}\log_e\left\{\epsilon + (1-\epsilon)e^{-\theta nr}\right\}. \tag{10}$$

Remark 2. *From* (10) *note that*

$$\lim_{\rho\to 0} R_E(\theta, n, r, \rho) = 0, \tag{11}$$

$$\lim_{\rho\to \infty} R_E(\theta, n, r, \rho) = r. \tag{12}$$

Note that the effective rate converges to zero in the low SNR regime, while it asymptotically approaches to the achievable rate r in the high SNR regime.

D. Traffic Model

The traffic arrival rate has a significant impact on the EEE, outage and delay of queueing systems [37], and it is hard to attain the real-world arrival processes. Therefore, We assume Markovian sources for simplicity and tractability of capturing arrival probability of the source, which is inherent characteristics to many IoT use cases [7]. We focus on the two-state ON-OFF discrete time Markovian sources model, in which the data arrival in the buffer is discrete in time [18].

In ON, state λ bits arrive in the buffer, whereas no data arrival occurs in OFF state as depicted in Fig. 1. The system has transition probability matrix $\mathrm{J} = (p)_{ij}$, where $p_{11} \in [0; 1]$ illustrates the probability of staying in OFF state, while $p_{22} \in [0; 1]$ determines the probability of ON state. The transition probabilities from one state to another are denoted by $p_{21} = 1 - p_{22}$ and $p_{12} = 1 - p_{11}$. In the steady state regime, the probability of ON state p_{on} is given as $p_{on} = \frac{1-p_{11}}{2-p_{11}-p_{22}}$ [31]. Thus, effective bandwidth of two-state discrete-time model characterized as

$$a^*(\theta) = \frac{1}{\theta}\log_e\left(\frac{\left(\phi + \sqrt{\phi^2 - 4(p_{11}+p_{22}-1)e^{\lambda\theta}}\right)}{2}\right) \tag{13}$$

$$\overset{(a)}{=} \frac{1}{\theta}\log_e\left(1 - p + pe^{\lambda\theta}\right), \tag{14}$$

where $\phi = p_{11} + p_{22}e^{\lambda\theta}$, and (a) is simplified version of effective bandwidth when we set $p_{11} = 1 - p$ and $p_{22} = p$, hence probability of ON state $p_{on} = p$ (refer to [31]). In (14), the single parameter p is considered as a measure of probability of arrival and it captures the variation of arrival rate from one instant to another.

We are interested in finding the average arrival rate of Markovian sources that can support the effective rate while satisfying the QoS constraints in (1). The QoS constraints are fulfilled when the effective bandwidth of the arrival process is equal to the effective rate of service process [31]; therefore,

$$a^*(\theta) = R_E(\theta, n, r, \rho). \tag{15}$$

To find the arrival rate that can support transmissions rate for given n, ρ and θ, we substitute the effective bandwidth expression of discrete-time Markov source (14) in (15) and solve for λ to obtain

$$\lambda = \frac{1}{\theta}\log_e\left(\frac{e^{\theta R_E(\theta,n,r,\rho)} - (1-p)}{p}\right). \tag{16}$$

Since $\lambda_{\text{avg}} = \lambda \cdot p$, which is equal to the average departure rate when the queue is in steady state; hence, we attain the average arrival rate as a function of QoS exponent θ, effective rate, and state transition probabilities as [31]

$$\lambda_{\text{avg}} = \frac{p}{\theta}\log_e\left(\frac{e^{\theta R_E(\theta,n,r,\rho)} - (1-p)}{p}\right). \tag{17}$$

III. Effective Energy Efficiency

From [27], EEE of the system is a ratio between R_E and total energy consumption at the transmitter under delay-outage probability constraint which can be measured in bits-per-joules/Hertz[4]. The energy consumption is mainly comprised of base-band processing blocks and radio-frequency (RF) chain. These blocks consist of low-noise amplifier (LNA), frequency synthesizers, digital-to-analog (AD/DA) and analog-to-digital converter, power amplifier (PA), filters and mixers. However, for a energy-limited IoT, the energy consumption of RF chain higher magnitude than base-band processing components. The power consumption of PA is directly proportional to the transmit power ρ and inverse drain efficiency ξ of the power

[4]When each channel-use is associated with a time-frequency unit, the effective rate under finite blocklength regime is measured in bits-per-second/Hertz and power in watts which can be written as Joules-per-second. Hence, EEE can be expressed as bits-per-joules/Hertz [38]–[40].

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.2022.3157956, IEEE Internet of Things Journal

amplifier (PA). Fore simplicity if base-band power consumption is neglected and the power consumption of all the other components in RF chain excluding PA is denoted P_c which is measured in watts. Then, simple power consumption model is defined as [27]

$$P_t = \xi\rho + P_c. \tag{18}$$

Thus, a simple performance benchmark for EEE of IoT devices analysis is given as

$$\mathrm{EEE} = -\frac{1}{n\theta}\frac{\log_e\left\{\epsilon + (1-\epsilon)e^{-\theta nr}\right\}}{\xi\rho + P_c} = \frac{R_E(\theta, n, r, \rho)}{P_t}. \tag{19}$$

Remark 3. *A well-known linear power consumption model has been extensively used in various studies regarding energy efficiency [41]. This model can capture the linear increment of power consumption, which occurs due to the enhancement of transmit power, circuit power, and inverse drain efficiency. It also assists with the analysis, which provides the performance benchmark for EEE in short packets and low latency communication. However, this model has the drawback of assuming the transmitter always has data for transmission.*

Remark 4. *In [21], the authors consider non-empty buffer probability for determining transmission mode. But this consideration holds in wireless fading channels when the source traffic is generated with a constant arrival rate. As a result, this model overestimate the EEE cant able to gives an accurate approximation of EEE.*

In (18), it is assumed that the buffer is always full during the transmission of a frame, hence the transmitter always in transmit mode which consumes more power. This consideration is unrealistic and overestimates energy consumption. In the formulation of EEE of wireless fading channels under random data arrivals and statistical queueing constraints, determining transmitter mode is a challenging task. Therefore, we explicitly consider two probabilistic events i.e., sources arrival and non-empty buffer, which attain the transmitter mode. Thus, modifying the basic power consumption model (18), by weighting the transmit mode with the transmission probability p_{ptx}, and idle mode with idle probability p_{idl} [28].

$$P_t = \xi\rho\ p_{\mathrm{ptx}} + \rho_{\mathrm{idl}} p_{\mathrm{idl}} + P_c, \tag{20}$$

where $p_{\mathrm{ptx}} = 1 - p_{\mathrm{idl}}$, then plugging it into (20), we have the obtained simplified power consumption model as follows

$$P_t = \xi\rho - (\xi\rho - \rho_{\mathrm{idl}})p_{\mathrm{idl}} + P_c. \tag{21}$$

The probability of the transmitter in an idle state depends upon the probabilities of two events. The First event that no data is generated at the source and the second event that the buffer is empty probability.

$$p_{\mathrm{idl}} = (1-p)(1-P_{nb}). \tag{22}$$

By substituting p_{idl} in (21) with (22), the power consumption mode can be expressed as

$$P_t = \xi\rho - (\xi\rho - \rho_{\mathrm{idl}})(1-p)(1-\frac{\lambda_{\mathrm{avg}}}{r}) + P_c. \tag{23}$$

Noted that $P_{nb} = 1$ or $p = 1$ holds the condition that buffer is always full. Hence, expressions (23) reduces to (18). Thereby, the corresponding new definition of the EEE

$$\mathrm{EEE} = -\frac{1}{n\theta}\frac{\log_e\left\{\epsilon + (1-\epsilon)e^{-\theta nr}\right\}}{\xi\rho - (\xi\rho - \rho_{\mathrm{idl}})(1-p)(1-\frac{\lambda_{\mathrm{avg}}}{r}) + P_c}. \tag{24}$$

A. Model Validation

Herein, we validate that considering the probability of a non-empty buffer in the power consumption model, which meets characteristics of the energy efficiency function of the Shannon's model. From [27], it is evident that with constant arrival rate, energy efficiency function must be non-negative or zero when the transmitting power approaches to zero, and this function tends to be zero as the transmitting power approaches to be ∞. We continue by confirming that the EEE in (24) is accurate even for the transmission of sporadic traffics.

Lemma 1. *The upper bound of EEE* (24) *is expressed as*

$$\mathrm{EEE}_{\infty} = \lim_{\rho\to\infty} \mathrm{EEE} = 0, \tag{25}$$

which describes that EEE is asymptotic at the high SNR towards 0. Meanwhile, there is a lower bound for EEE, which approaches zero for very low power as follows

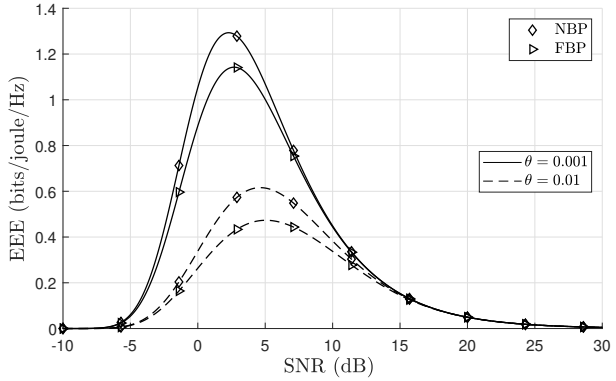
$$\mathrm{EEE}_0 = \lim_{\rho\to 0} \mathrm{EEE} = 0, \tag{26}$$

Proof. The proof can be found in Appendix A. □

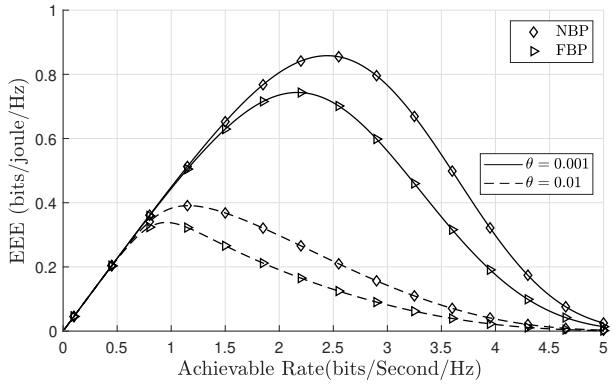
Next, we investigate the dependence of EEE on ρ and r and analyze the behaviour of EEE by varying these two parameters. Fig. 2. (a) depicts the EEE as a function of ρ while Fig. 2. (b) shows the EEE as a function of r, for fixed values of m, n and different delay exponents θ. It is observed that the EEE curves decline when a stringent QoS constraint $(\theta \to \infty)$ is imposed. The main intuition of Fig. 2. (a) and Fig. 2. (b) is that the EEE is likely a quasi-concave function with respect to r and ρ [23]. Taking one step further, the quasi-concavity can be inferred with the help of **Lemma 1** which indicates the at the bounds, the EEE tends to zero with respect to ρ [21]. Moreover, it is also known that the effective rate and the power consumption are both differentiable with respect to r and ρ, while the denominator in the EEE is strictly increasing in ρ, and decreasing in r, which are strong indicators of quasi-concavity. The latter is a consequence of infrequent packet transmission, which is a more realistic approach for modelling IoT traffic (due to the inherent characteristics of short packets with periodic or sporadic arrival) and energy efficiency [7]. The full buffer model overestimates the energy consumption, while the non-empty buffer probability model captures the sporadic nature of the transmissions. More importantly, efficient use of r and ρ boosts the performance of communication system in terms of energy efficiency, which prompts us to seek the pair (r^*, ρ^*) that maximizes the EEE as we shall discuss in the next section.

Remark 5. *Note that* (15) *circumvent data losses and ensure system stability while maximizing the EEE. Hence* $R_E(\theta, n, r, \rho) \geq \lambda_{\mathrm{avg}}$ *is always true. Furthermore, the* λ_{avg}

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.2022.3157956, IEEE Internet of Things Journal



(a)



(b)

Fig. 2. (a) EEE as a function of SNR for fixed vales of $r = 1$, $n = 500$, $p = 0.5$, $m = 2$, $\xi = 0.2$, $P_c = 0.2$ watts and $p_{\text{idl}} = 0.03$ watts. (b) EEE as a function of rate for fixed vales of SNR $= 10$dB, $n = 500$, $p = 0.5$, $m = 2$, $\xi = 0.2$, $P_c = 0.2$ watts and $p_{\text{idl}} = 0.03$ watts.

increases with p, which we will see later in Fig. 7. *Lower values of p indicate that the data arrives less frequently, whereas, $p = 1$, represents that the source is always in ON state and average arrival rate λ_{avg} equal to the effective rate [31].*

IV. Maximization of Effective Energy Efficiency with QoS Guarantees

The efficient utilization of radio resources such as transmission power and rate need to be optimized to ensure the QoS guarantees and EEE in the IoT network. Thereby, we emphasis on rate and power allocation optimization strategies, as those resources generally form refined radio management schemes. Note that power allocation strategies have hardware limitations since there are maximum and minimum power levels and transmit hardware can manipulate them. Besides that, when dealing with various applications, in the sense that devices can carry data over variable data rates, the transmission rate is a degree of freedom (DoF) that can be utilized. Such types of resource allocation strategies have been exploited in different setups such as power allocation is mainly used to increase the battery life span in LPWAN (low power wide area network) INGENU technologies [42], while LoRa supports rate control often associated with fixed transmission power levels [43].

In order to find optimal resource allocation strategies for the proposed scenario, we formulate different EEE maximization problem constrained on effective rate along with minimal transmission power allocation (PA), maximal rate allocation (RA) or joint PA and RA resource allocation strategies, respectively. Moreover, transmission power is upper bounded by P_{max}.

Hence, the optimization problems can be formulated as $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$. We first formulate PA problem as $\mathbf{P1}$ in which we first find the optimum power ρ^* that satisfies QoS constraints and gives maximum EEE for a fixed r.

$$\mathbf{P1}: \quad \text{maximize}_{\rho \in \mathcal{S}} \quad \text{EEE}, \tag{27}$$

$$\text{subject to} \quad \rho \leq P_{max}. \tag{28}$$

where $\mathcal{S} \subseteq \mathbb{R}_+$, EEE, $R_E(\theta, n, r, \rho) : \mathcal{S} \rightarrow \mathbb{R}_+$, EEE ≥ 0, $R_E(\theta, n, r, \rho) \geq 0$.

We now reformulate RA problem as $\mathbf{P2}$ in which optimum rate r^* attains maximum the EEE for a fixed ρ in presence of QoS constraints.

$$\mathbf{P2}: \quad \text{maximize}_{r \in \mathcal{S}} \quad \text{EEE}, \tag{29}$$

$$\text{subject to} \quad \rho \leq P_{max}. \tag{30}$$

We finally propose an iterative algorithm to control transmit powers and rates (PA and RA) jointly as problem $\mathbf{P3}$ to maximize the EEE under QoS constraints based on the results in previous problems ($\mathbf{P1}$ and $\mathbf{P2}$).

$$\mathbf{P3}: \quad \text{maximize}_{\rho \in \mathcal{S}, r \in \mathcal{S}} \quad \text{EEE}, \tag{31}$$

$$\text{subject to} \quad \rho \leq P_{max}. \tag{32}$$

Due to the complexity in the EEE expression, the optimization problem is quite challenging to obtain a closed-form solution. Note that $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$ are composed of a ratio between two functions; R_E and power consumption P_t; both of these are the functions of r and ρ. These forms of optimization problems are widely solved by using fractional programming, particularly the popular Dinkelbach's (DK) algorithm, or alternatively via Cross Entropy (CE) Optimization, as compared to the time-consuming exhaustive searching algorithms [44]–[46].

A. Modified Dinkelbach Method

For completeness, in what follows we describe the key components of fractional programming and then apply it to problems $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$. The general non-linear form of fractional programming is

$$\text{maximize}_{x \in \mathcal{S}} \quad q(x) = \frac{f_1(x)}{f_2(x)}, \tag{33}$$

where $\mathcal{S} \subseteq \mathbb{R}^n, f_1, f_2 : \mathcal{S} \rightarrow \mathbb{R}, f_1(x) \geq 0$ and $f_2(x) > 0$. Using parametric convex program [44], the objective of the fractional programming problem in (33) is first transformed into an equivalent fractional program as

$$\begin{aligned} &\text{maximize}_{x \in \mathcal{S}, \Lambda \in \mathbb{R}} \quad \Lambda \\ &\text{subject to} \quad f_1(x) - \Lambda f_2(x) \geq 0. \end{aligned} \tag{34}$$

Now, the function is expressed as

$$F(\Lambda) = \text{maximize}_{x \in \mathcal{S}} \quad f_1(x) - \Lambda f_2(x). \tag{35}$$

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.2022.3157956, IEEE Internet of Things Journal

Note that $F(\Lambda)$ is convex, continuous and strictly decreasing with Λ [44], and that (35) can be observed as a scalarized bi-criterion optimization problem. $f_1(x)$ is to be maximized and $f_2(x)$ is to be minimized, and Λ controls the relative weight of the denominator. Thus, the following statements are equivalent $F(\Lambda) = 0 \iff \Lambda = q^*$, where q^* is the optimal value of the (33) which is attained when $F(\Lambda)$ converges to zero. The optimum solution of (33) is equivalent to obtain the root of the non-linear function $F(\Lambda)$. Hence, the optimally condition is $F(\Lambda^*) = \text{maximize}_{x\in\mathcal{S}} \quad f_1(x) - \Lambda^* f_2(x) = 0$. In [47] an appealing example for the optimal q_n^* for a given value of Λ_n, which is based on Newton's method and expressed as follows:

$$\Lambda_{n+1} = \Lambda_n - \frac{F(\Lambda_n)}{F'(\Lambda_n)} = \Lambda_n - \frac{f_1(x_n^*) - \Lambda_n f_2(x_n^*)}{-f_2(x_n^*)} = \frac{f_1(x_n^*)}{f_2(x_n^*)}, \tag{36}$$

where x_n^* is optimal for a given value of Λ.

After applying fractional programming technique to our optimization problems, **P1**, **P2** and **P3**. They can be re-written by using a parametric convex program **P4** as

$$\mathbf{P4}: \quad f(x,q) = \text{maximize}_{x\in\mathcal{S}} \quad f_1(x) - q f_2(x). \tag{37}$$

$$q_{n+1} = \frac{f_1(x)}{f_2(x)}, \tag{38}$$

where $f_1(x) = R_E$, $f_2(x) = P_t$ and x is the optimization variable that represents ρ in **P1**, r in **P2**, and ρ and r in **P3**. Dinkelbach's method relies on solving inner loop optimization as **P4** in each iteration [44]. Due to the simplicity and improving the convergence of Dinkelbach's, Golden Section Search method is consider to solve (37) problem.

The intuition behind the modified Dinkelbach's **Algorithm 1** is as follows. It begins from a certain arbitrary estimate parameter of q and evaluates the level set of the problem defined in (37) by Golden Section Search method. Note that the upper and lower bound vectors are considered on [ρ] for solving **P1**, [r] in **P2**, and [ρ r] in **P3**. With initial values of bounds vector, ub being the upper bound P_{max} and lb as 0.001, golden Section Search method finds the maximum of function (37) by narrowing intermediate bounds vectors until stopping criteria is satisfied. Then, as Dinkelbach starts to converge, it adjusts the estimated value of q iteratively via (38) and checks whether the stopping criteria is achieved or not; herein we have set a maximum number of iterations. If the tolerance gap still exits, then the algorithm proceeds to search over the level set function (37) until it settles within the tolerance margins[5]. Finally, the algorithm converges to a optimum solution.

Algorithm 1 Modified Dinkelbach's (DK) Algorithm

Initialization : lb, ub are the lower and upper-bounds, error, tolerance parameters $\tau_1 = 10^{-6}$,$\tau_2 = 10^{-6}$, $q \leftarrow 0$, $i \leftarrow 0$, optimized $\leftarrow$ false
Define :$f(x,q) = \mathbf{P4}$
while (optimized $\leftarrow$ false and max-iterations) **do**
  Compute $\eta_1 = ub - (ub - lb) * 0.618$
  Compute $\eta_2 = lb + (ub - lb) * 0.618$
  Compute $g_1 = f(\eta_1, q)$
  Compute $g_2 = f(\eta_2, q)$
  while (error $\geq \tau_1$) **do**
    if $g_1 > g_2$ **then**
      $ub \leftarrow \eta_2; \eta_2 \leftarrow \eta_1; g_2 \leftarrow g_1$
      Compute $\eta_1 = ub - (ub - lb) * 0.618$
      Compute $g_1 = f(\eta_1, q)$
    else if $g_1 < g_2$ **then**
      $lb \leftarrow \eta_1; \eta_1 \leftarrow \eta_2; g_1 \leftarrow g_2$
      Compute $\eta_2 = lb - (ub - lb) * 0.618$
      Compute $g_2 = f(\eta_2, q)$
    end if
    Compute error $= 2 * \frac{ub-lb}{ub+lb}$
  end while
  Compute $x = \frac{\eta_1+\eta_2}{2}$ and reset lb, ub
  if $f(x,q) = 0$ **then**
    $x^* \leftarrow x$
    optimized $\leftarrow$ true
  else if $f(x,q) \leq \tau_2$ **then**
    $x^* \leftarrow x$
    optimized $\leftarrow$ true
  else
    update, $q \leftarrow \frac{f_1(x)}{f_2(x)}$
    update, $i \leftarrow i + 1$
  end if
end while

B. Cross Entropy-based Solution

The cross entropy (CE) [46], [48] in a **Algorithm 2** is also able to solve the optimization problems **P1**, **P2** and **P3** by obtaining the corresponding optimal points expressed as

$$\gamma* = P = \max_{X\in\Omega} \quad P(X), \tag{39}$$

where X is a generic point in sample space Ω and P denotes the performance function (objective function **P1**, **P2** and **P3**). In the next phase, the CE framework describes the associated stochastic problem. Therefore, (40) is revised based on the estimation of the probability $Pu\{.\}$ for a fixed parameter vector $u \in \psi$ i.e,

$$l(\gamma) = P_u\{P(X) \geq \gamma\} = \mathbb{E}\ \mathbb{I}_{P(X)\geq\gamma}, \tag{40}$$

where X is a random set of samples that are generated by PDF $f(.;u)$ which belongs to the family of PDF $\{f(.;v)v \in \psi\}$ from the discrete space set Ω, $\mathbb{E}$ is the expectation operator and $\mathbb{I}\{.\}$ denotes the indicator function of event$\{.\}$. This is an iterative method dealing with two tasks: *i)* generates random samples according to a specified mechanism, and *ii)*, the parameters of the random mechanism are updated at each iteration based on this data to produce an updated solution.

The main intuition behind the Cross Entropy **Algorithm 2** is that it begins to generate a set of random samples $X_1, X_2, ..., X_k$ using distribution function $f(,;v)$ with mean μ_0 and variance σ_0^2 i.e $v = [\mu_0, \sigma_0^2]$ where σ_0^2 is the upper bound with initial values of P_{max} and then it determines the threshold γ as the $(1-z) - quantile$ from the performance

[5]The tolerance parameters in **Algorithm 1** are set to $\tau_1 = 10^{-6}$ and $\tau_2 = 10^{-6}$ for Golden Search and Dinkelbach's Algorithm respectively.

Algorithm 2 Cross Entropy (CE) Algorithm

1: **Initialization** : μ_0 and ${\sigma_0}^2$, number of samples N, threshold γ, rarity parameter d, tolerance τ, number of iterations itr_n and maximum number of iterations max_n
2: Set $k = 0$
3: **while** ($itr_n \leq \ max_n$) **do**
4: Generate N number of samples $X_1, ..., X_N$ from sampling distribution $\mathcal{N}(\mu_k, {\sigma_k}^2)$
5: Compute performance values as $P(X_1), ..., P(X_N)$
6: Compute threshold as $(1-d)quantile = \gamma_k$ from performance values of the sample $P(X_1), ..., P(X_N)$
7: Compute $Elite$ samples vector, i.e., the samples that satisfied $P(X_N) \geq \gamma_k$
8: Count elements of $Elite$ samples vector $= C$
9: Update $\mu_{k+1} = \frac{1}{C}\sum_{i=1}^{C} Elite_i$
10: Update $\sigma^2_{k+1} = \frac{1}{C-1}\sum_{i=1}^{C}(Elite_i - \mu_{k+1})^2$
11: **if** $(\mu_{k+1} - \mu_k) \geq 10^{-6} \&\& (\sigma^2_{k+1} - \sigma^2_k) \geq 10^{-6}$ **then**
12: Break
13: **end if**
14: $itr_n = itr_n + 1$
15: **end while**

values $P(X_1), ..., P(X_k)$, where z is generally denoted as the rarity parameter [48]. The elite samples can be obtained from those samples whose performance is better than threshold γ as $P(X_k) \geq \gamma$. This forms new parameterized distribution vector $f(.; v')$ which is estimated to be close at the target distribution $f(.; v^*)$ by minimizing the Kullback-Leibler divergence, i.e., cross-entropy. This step executes a single iteration. Based on these steps, CE algorithm iteratively updates $f(.; v)$ to generate a family of PDFs $f(.; v_1), f(.; v_2), ..., f(.; v^*)$ with the help of threshold values $\gamma_1, \gamma_2, ..., \gamma^*$ to obtain the optimal density function$f(.; v^*)$. This procedure shows that the CE algorithm is easy to implement and effective to converge to the optimal solution.

C. Computational Complexity Analysis

In order to evaluate the computational complexity of the proposed algorithms, we consider the worst-case complexity using big-$\mathcal{O}$ notation, which is commonly adopted in several resource allocation studies [49]. As a result, **Algorithm 1** is composed of two nested loops: outer loop **L2** and inner loop **L1**. **L2** is used to calculate better response of (38) and (37) at each iteration, while **L1** is used to update q according to the golden search algorithm. For the optimal convergence of algorithm, the complexity is dominated by **L2** which has computation complexity equal to $\mathcal{O}(k \log_2(\frac{ub-lb}{\tau_2}))$, where k is number of iteration required for optimum values [50]. For maximum EEE, **L1** could find the zero of $f(x, q)$ by updating q according to the golden search method which gives us a fair estimate, and can be found within tolerance τ_1, whose computation complexity becomes $\mathcal{O}(\log(\frac{1}{\tau_1}))$ iterations [49]. Finally, we can obtain an overall asymptotic complexity of $\mathcal{O}(k(\log(\frac{1}{\tau_1}))(\log_2(\frac{ub-lb}{\tau_2}))$.

Complexity analysis of traditional CE **Algorithm 2** has been initiated in [51], where authors proved a ground-breaking stochastic runtime result for the CE. From **Algorithm 2** the runtime results proved that sample and elite size plays a crucial role in efficiently finding an optimal solution. It can clearly observe that mainly complexity comes from step 5 to 7, that showed if the elite vector size C is moderately adapted to the sample size N, then the stochastic runtime of the traditional CE is $\mathcal{O}(CN)$ for arbitrarily small $\tau > 0$ and a constant smoothing parameter $d > 0$. This can be also expressed as $\mathcal{O}(k^3)$ because these three operation mainly depend upon iteratively operation. Lastly, it is proved that the computation complexity of the **Algorithm 1** $\mathcal{O}(k(\log(\frac{1}{\tau_1}))(\log_2(\frac{ub-lb}{\tau_2}))$ typically much smaller than **Algorithm 2** [52].

V. Results and Discussion

In this section, we present numerical results for the performance of the optimal resource allocation strategies. First, we confirming that the EEE in (24) is accurate model even for the transmission of sporadic traffics. Further, we evaluate the convergence performance of different algorithms for the resource allocation strategies till attain the optimum EEE. We also highlight that constant arrival data increases the λ_{avg} which consumes more energy as compared to a lower arrival rate. Finally, we show the trade-off between the EEE, outage-delay violation probability and efficient utilization of the resource allocation algorithms. For the following results, we fix the network parameters as follows P_c and p_{idl} are set to 0.2 and 0.03 watts, $\xi = 0.2$, $d = 500$, $m = 2$, $n = 500$, while solving $\mathbf{P1}$ and $\mathbf{P2}$, we assume $r = 1$ bps/Hz and $\rho = 10$ dB respectively.

Next, we evaluate the different EEE models for different transmission probabilities (i.e, when the buffer is not empty) as illustrated in Fig. 3. It is expected that the EEE monotonically decreases with the increase of arrival rate which indicates higher congestion in the network that can be steered by different values of p. The results from the analytical model as in (24), demonstrate that this model is much more accurate as compared to the analytical model of energy efficiency (19) and [21, Eq. (19)]. This is because it captures the sporadic traffics behaviour of IoT devices. For example when $p = 1$, it indicates constant arrival in the buffer. Under such circumstances, the transmitter is always busy (i.e., $p_{\text{idl}} = 0$), it is expected that EEE models efficiency values become identical to FBP. In this case, the model in [21, Eq. (19)] performs as an upper bound.

A. Convergence of the Proposed Algorithms

Next, we compare the proposed resource allocation strategies evaluated with the proposed algorithms and Matlab *fmincom* for comparison. We will see that joint PA and RA outperforms significantly as compared to the other two approaches.

We demonstrate the convergence performance of different algorithms for resource allocation strategies. The obtained optimum EEE at each different iteration rounds is illustrated in Fig. 4 and the corresponding main results such as average simulation execution time[6], optimum power and rate are

[6]Evaluation of execution time on an IntelCore i5-8250 CPU@1.60 GHz processor with 16 GB RAM.

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.2022.3157956, IEEE Internet of Things Journal

TABLE II
ALGORITHMS COMPARISON RESULTS OF THE DIFFERENT EEE FORMULATION .

Parameters	RA			PA			Joint PA and RA		
	CE	**DK**	$fmincon$	**CE**	**DK**	$fmincon$	**CE**	**DK**	$fmincon$
iteration	07	**02**	07	10	**06**	09	11	**05**	20
execution time (s)	0.2525	**0.0310**	1.2783	0.6430	**0.0604**	1.2810	0.6496	**0.2198**	1.1198
Power*	-	-	-	1.6700	1.6922	1.6922	2.2311	2.6478	1.8318
Rate*	2.166	2.167	2.167	-	-	-	1.1181	1.4217	1.4320
EEE*	0.7435	0.7435	0.7435	1.2931	1.2932	1.2932	1.0817	1.6416	1.1198
RE	1.6357	1.6357	1.6357	0.6062	0.6120	0.6120	0.7346	0.7864	0.5243
NBP	0.7552	0.7552	0.7552	0.6060	0.6120	0.6120	0.6568	0.5531	0.3660

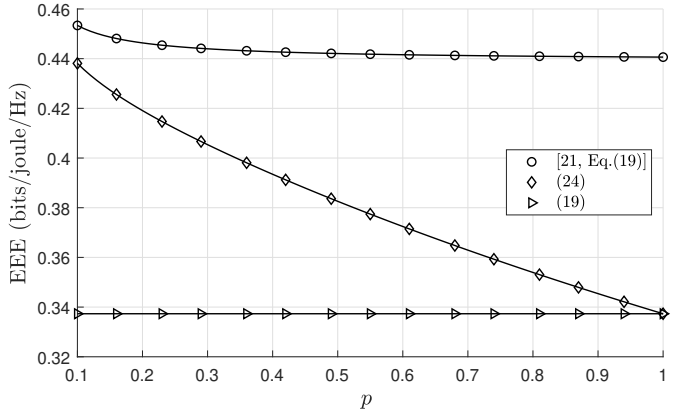


Fig. 3. EEE as a function of p for fixed vales of $r = 1$, $\rho = 10$ db, $\theta = 0.01$, $n = 500$, $m = 2$, $\xi = 0.2$, $P_c = 0.2$ watts and $p_{\text{idl}} = 0.03$ watts.

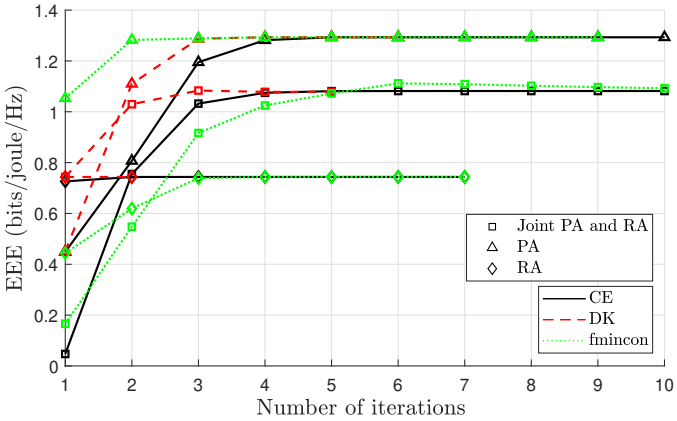


Fig. 4. Convergence performance of different resource allocation strategies with distinct algorithms with $m = 2, n = 500, \theta = 0.001$ and $p = 0.5$ for Dinkelbach's, Cross Entropy and interior-point ($fmincon$) algorithm.

shown in Table II. Here, we consider network parameters as $\theta = 0.001, m = 2, n = 500$ and $p = 0.5$. It is observed that DK converged to similar optimum power, rate and EEE values as CE and interior point ($fmincon$) algorithms which can be also seen in Table II. That ensured the accuracy of algorithms. However, DK converged faster than others because it integrates with the state-of-the-art Golden Section Search algorithm for reducing computation complexity as discuss in Section IV-C. It reduced the amount of iterations leading to the usage fewer resources as it does not rely on random sampling in each iteration as CE. In other words, to attain the optimal EEE in joint PA and RA, DK transforms 5 iterations and takes 0.2198 seconds, while the CE transforms 11 iterations within 0.6496 seconds. In PA resources allocation, as CE executes 10 iteration rounds in 0.6430 seconds, whereas, DK executes 6 iterations in 0.0604 seconds. For RA resources allocation, DK requires almost the similar number of iterations, though requires fewer resources. Therefore, the above numerical results verify that for joint PA and RA optimization, the DK's algorithm has high performance and low complexity as compared to the CE and interior point ($fmincon$) algorithms.

Therefore, in the remaining of this section we use the proposed modified DK algorithm, due to its efficiency in convergence speed and use of computational resources.

B. Optimal EEE under Markovian Arrivals

Fig. 5 shows the achieved maximum EEE obtained from **P1**, **P2** and **P3** for different n with for fixed values of m and p. It is observed that a lower value of n improves the EEE, due to the fewer channel resources utilized for transmission of packets. The looser delay constraints reduce the possibility of buffer congestion consequently, which also leads to the higher EEE as compared to the stringent QoS constraints. Furthermore, it is concluded that joint rate and power optimization outperforms compared to sole rate or power optimization in loose delay requirements, due to more degrees of freedom in optimal selection of rate and power. On the other hand, when tight QoS constraints are imposed, different resource allocation strategies present significantly less performance gap, due to strict QoS imposition (large θ) which restricts the solution domain.

Fig. 6 shows the EEE as a function of n. Notice that a higher value of n reduces the EEE. Moreover, higher λ_{avg} (higher p) induce a decrease in the EEE, owing to the higher P_{nb} values. Moreover, it is notice that Joint PA and RA is the optimal power allocation strategy as compared to others PA or RA in presence of constant or random arrival traffics.

Fig. 7. (a) illustrates the λ_{avg} in (17) as a function of p for fixed values of QoS constraint θ, m, and n. It is observed that lower values of p decrease λ_{avg} and different values

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.2022.3157956, IEEE Internet of Things Journal

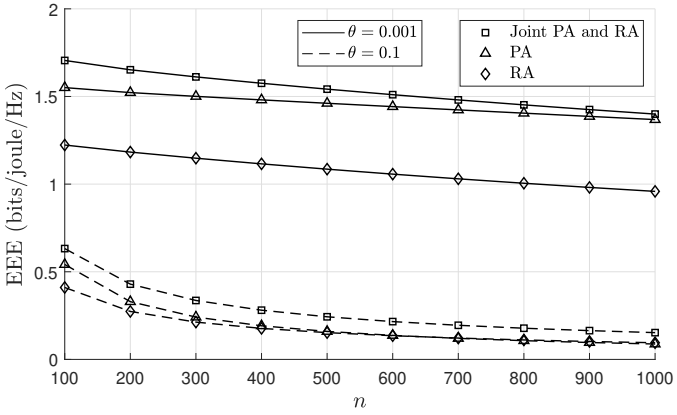


Fig. 5. EEE as a function of n for fixed values of $d = 500$, $m = 2$ and $p = 0.1$.

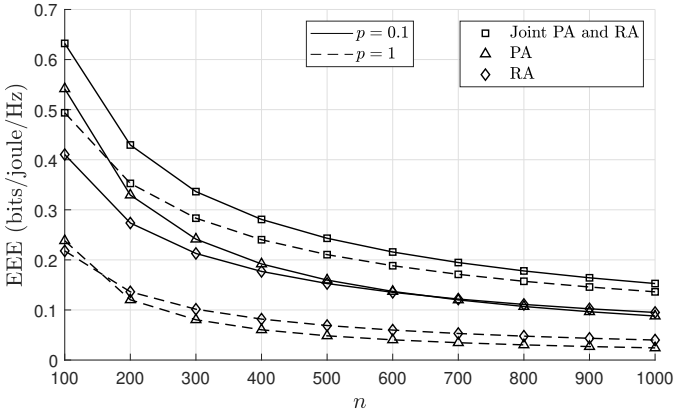
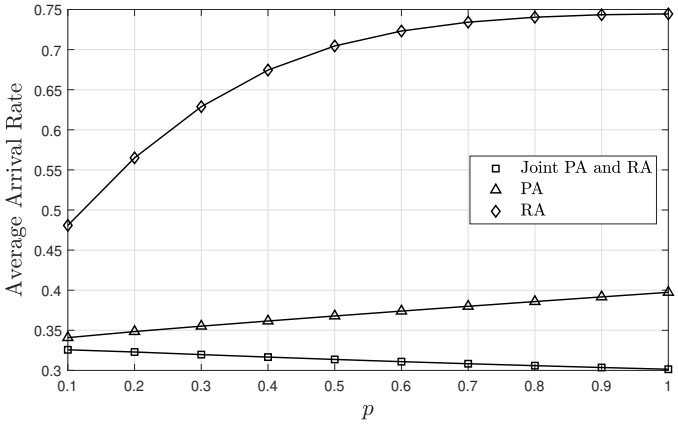
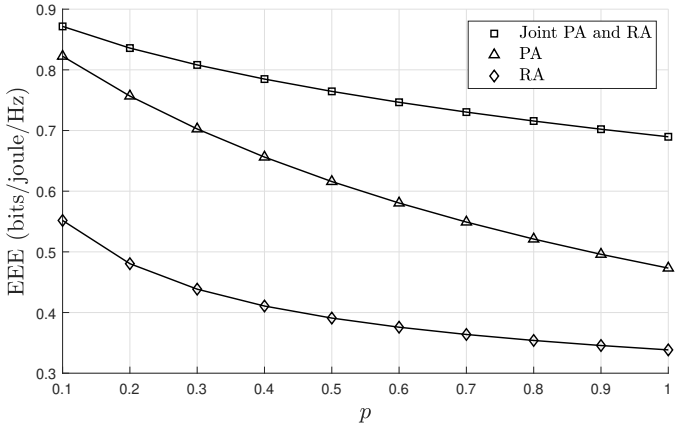


Fig. 6. EEE as a function of n for fixed values of $\theta = 0.1$, $d = 500$ and $m = 2$.



(a)



(b)

Fig. 7. (a) Average arrival rate as a function of p for fixed values of $\theta = 0.1$, $n = 500$, $d = 500$ and $m = 2$. (b) EEE as a function of p for fixed values of $\theta = 0.1$, $n = 500$, $d = 500$ and $m = 2$.

p captures the constant and sporadic traffics behaviour of IoT devices. However, this is a simplification of the arrival process to facilitate the analytical tractability and provide some insights on the interplay of average arrival and effective rates. The variation of the λ_{avg} is characterized by the p parameter. Therefore, each variation in λ_{avg} approaches distinct optimum power and rate values in optimization problems, defined in $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$. To analyze the impact of λ_{avg} on the EEE, substitute these optimum values in (17). Thus, from the Fig. 7. (b) it is observed that higher traffic arrival consumes more energy as compared to a lower arrival rate because there are more packets to transmit. Furthermore, joint PA and RA allocation performs better than the alternatives when faced with sporadic traffic.

Fig. 8. (a) shows the delay violation probability and Fig. 8. (b) outage probability as a function of n for fixed values of QoS constraint θ, m and p. The optimum power and rate values, obtained from solving optimization problems $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$ w.r.t n. It is conclude from earlier Fig. 5 and 8. (a) that joint PA and RA is optimal power allocation strategy with higher outage and delay violation probability as compared to PA and RA strategies. This is because of our assumption, (i.e. $\rho = 10$ dB) RA strategy always gives higher values for optimal r and smaller P_{nb} as compared to the joint PA and RA gives as illustrated in Table II. Thus, it decreases delay violation probability significantly (delay violation probability function is inversely proportional to r as in (3)). Furthermore, larger values of n increase the delay violation probability and decrease the outage probability. Short packets (lower n) have low delay violation probability because of short waiting time in the buffer. Consequently, they hardly congest the buffer and thus satisfy the latency requirement. On the other hand, short packets are susceptible to fading which induces decoding errors at receiver. Moreover, average arrival rate in the ON state is λ_{avg}, since we fix such probability and the equality between effective bandwidth and rate, the source adapts the rate. In this case, higher values of p induce an increase in the λ_{avg}, and in turn the source is perceived as constant arrival, which directly influences the reliability and latency metrics of the system. Note that since maximizing the EEE does not maximizes the effective rate, RA or PA may perform better than the joint allocation because the energy efficiency is much stricter in such case. However, this occurs due to high energy consumption, which is not feasible for energy-limited systems such as massive IoT or MTC deployments.

Fig. 9. (a) shows the delay violation and Fig. 9. (b) outage probability as a function of fading parameter m. We note that both delay violation and outage probability decrease sharply

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.2022.3157956, IEEE Internet of Things Journal

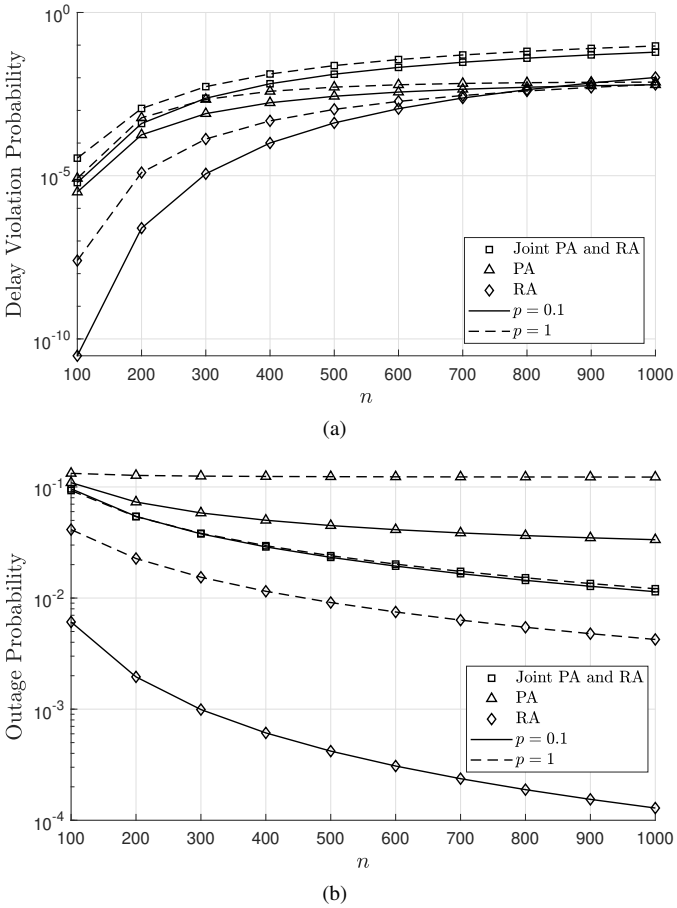


Fig. 8. (a) Delay violation probability as a function n for fixed values of $d = 500$, $m = 2$ and $\theta = 0.1$. (b) Outage probability as a function n for fixed values of $d = 500$, $m = 2$ and $\theta = 0.1$.

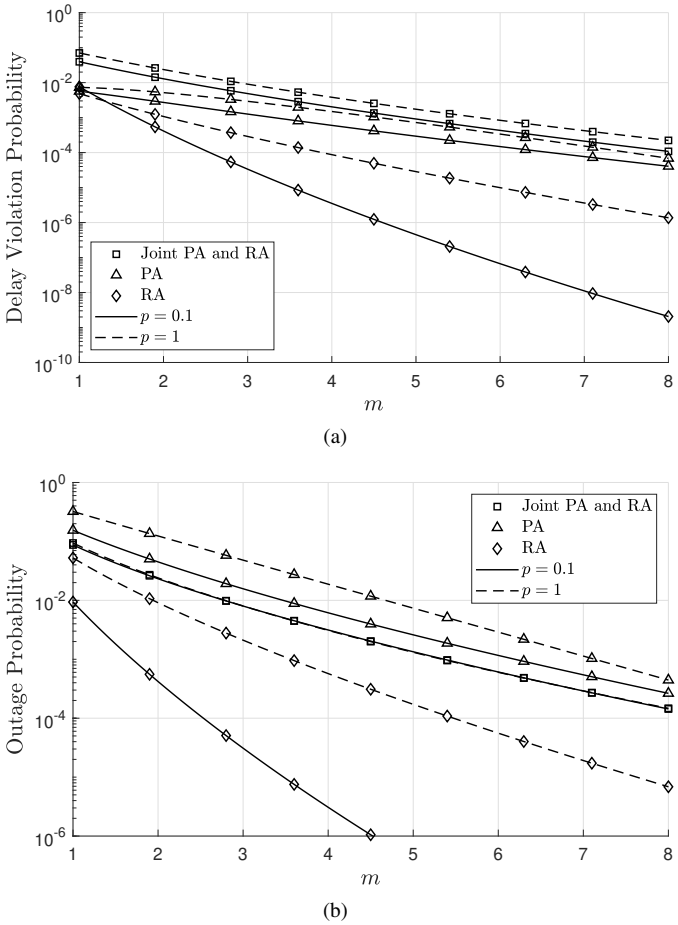


Fig. 9. (a) Delay violation probability as a function m for fixed values of $d = 500$, $n = 500$ and $\theta = 0.1$. (b) Outage probability as a function m for fixed values of $d = 500$, $n = 500$ and $\theta = 0.1$.

with the increase of the value of m. The reason is that the channels experience less uncertainty due to the increased LoS component at the receiver, turning the channel more deterministic and inducing lower delay. Notice that there is a significant gap between constant and sporadic arrival traffic because constant sources may have higher arrival rate which yields larger delay violation and outage probability, since queues build up quickly with larger λ_{avg}, while the service is limited by n. Both figures depict that although joint PA and RA is the optimal EEE strategy, but it renders higher outage and delay violation probability. Thus, RA should be used for those IoT applications which demand minimum outage and minimum delay violation at cost of reduced EEE.

The impact of θ is examined in Fig. 10. (a) and Fig. 10. (b), where we plot the outage and delay violation probabilities with respect to QoS exponent θ. We solve the optimization problems $\mathbf{P1}$, $\mathbf{P2}$ and $\mathbf{P3}$ numerically for every value of θ. It shows that PA or RA resource allocation strategy renders lower delay and outage probabilities for higher values of θ. Moreover, examining the figure, it is evident that constant arrival when $p = 1$ also increase delay violation and outage probabilities due to higher arrival rate and P_{nb}.

Fig. 11 illustrates the delay violation probability for different delay bounds d. Larger values of delay bound d imply that the system can support longer delay. Thus, delay violation probability decreases as d increases. Furthermore, it is evinced that, as aforementioned, constant arrival when $p = 1$ consume more energy and has higher arrival probability, which also increase delay violation probability due to the buffer congestion. Similar to the Fig. 8, that joint PA and RA optimization has higher delay violation probability, while rate allocation results in smaller delay violation probability because of the high fixed value of ρ in RA, which is also confirmed from **Lemma 1**.

VI. Conclusion

We have studied energy efficient power and rate allocation schemes of power-limited IoT systems operating under statistical QoS constraints. We resorted to a non-empty buffer probability-based energy consumption model that also accounts for short packets. Markovian arrivals allowed us to capture some key characteristics of IoT devices traffic such as constant and sporadic traffic arrival processes. We formulated optimization problems for maximizing the EEE while satisfying the required QoS constraints imposed by Markovian arrival. The solution determined the optimal power and rate levels via PA, RA, and joint PA and RA optimizations. We used different state-of-the-art algorithms such as DK, CE, and interior point to the obtain optimal solution of such non-convex problem and compared the performance of each

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.2022.3157956, IEEE Internet of Things Journal

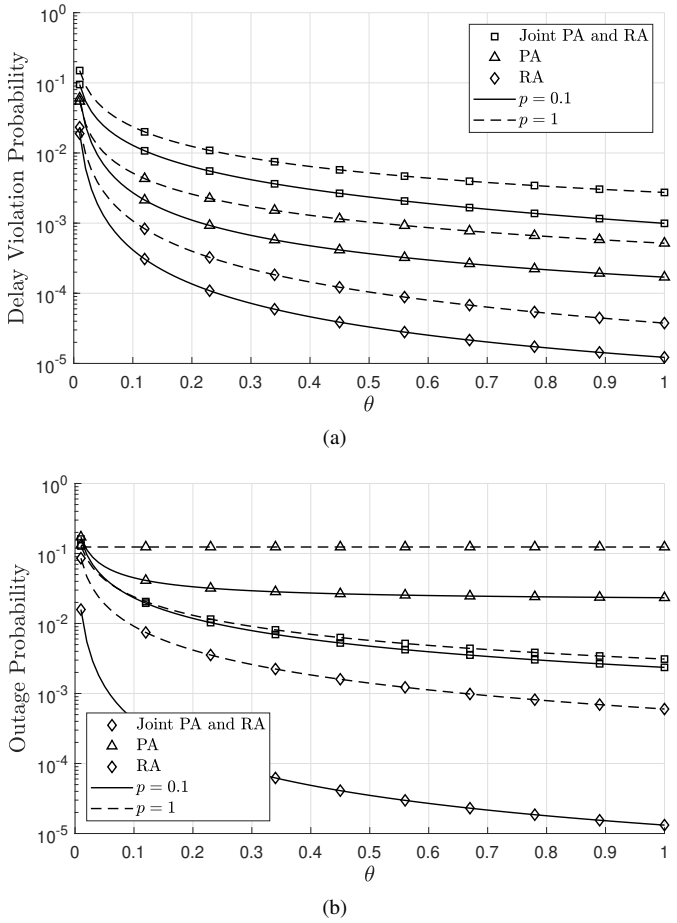


Fig. 10. (a) Delay violation probability as a function θ for fixed values of $d = 500$, $n = 500$ and $m = 2$. (b) Outage probability as a function θ for fixed values of $d = 500$, $n = 500$ and $m = 2$.

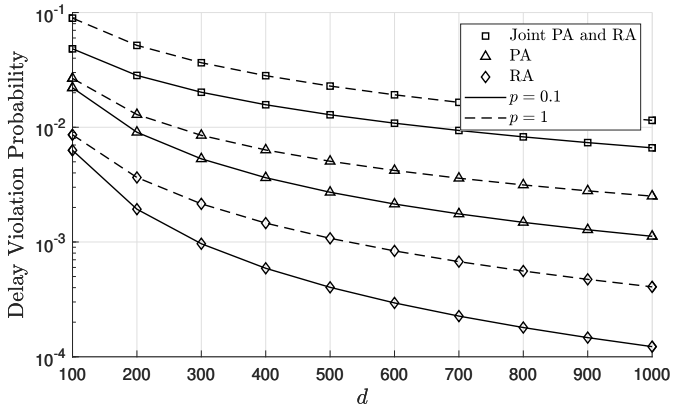


Fig. 11. Delay violation as a function d for fixed values of $n = 500$, $m = 2$ and $\theta = 0.1$.

iterative algorithms which includes guarantees to convergence to the optimal solution.

Our results have drawn some key design intuitions on the impact of traffic on the performance of the communication system. We show that joint PA and RA is optimal resource allocation strategy but have higher outage and delay violation probability because higher fixed transmission power in RA or transmission rate in PA maximize effective rate but not necessarily the EEE. Therefore, these resource allocation schemes support lower outage and delay violation probability while compromising on EEE. The substantial performance gain is achieved through the significance of LoS component for power-constrained devices. The energy consumption model based on the non-empty buffer probability leverages the proposed power and rate allocation strategy. Consequently, the resource scarcity is improved at the source by adjusting the arrival rate. We have further investigated the impact of the delay violation and outage probability, transmission power, rate, QoS constraints, constant and sporadic arrival traffics, and blocklength on the EEE. Lastly, our proposed hybrid algorithm, Dinkelbach's with Golden Section Search, rendered higher convergence performance than other algorithms.

APPENDIX A
PROOF OF LEMMA1

From (8), the outage probability can be expressed with $\vartheta = \alpha + \sqrt{\frac{\pi}{2}\mu^{-2}}$ and $\varrho = \alpha - \sqrt{\frac{\pi}{2}\mu^{-2}}$. The, by substituting the values of α and μ, we have

$$\vartheta = \frac{1}{\rho}\left\{2^r - 1 + \sqrt{\left(\frac{\pi^2(2^{2r}-1)}{n}\right)}\right\}. \tag{41}$$

Similarly, ϱ is

$$\varrho = \frac{1}{\rho}\left\{2^r - 1 - \sqrt{\left(\frac{\pi^2(2^{2r}-1)}{n}\right)}\right\}. \tag{42}$$

Thus, if $\rho \to \infty$ then $\vartheta, \varrho = 0$ and $\Gamma(m, 0) = 0$, while on the other end, if $\rho \to 0$ then $\vartheta, \varrho = \infty$ and $\Gamma(m, \infty) = 1$. Hence, substituting the $\Gamma(m, 0)$ and $\Gamma(m, \infty)$ into (8), outage probability bound at $\lim_{\rho\to 0} \epsilon = 1$ and $\lim_{\rho\to\infty} \epsilon = 0$. Then,

$$\lim_{\rho\to 0} C_E(\theta, n, r, \rho) = -\frac{1}{n\theta}\log_e\left\{\epsilon + (1-\epsilon)e^{-\theta n r}\right\} = 0.$$

$$\lim_{\rho\to\infty} C_E(\theta, n, r, \rho) = -\frac{1}{n\theta}\log_e\left\{0 + (1-0)e^{-\theta n r}\right\} = r.$$

Based on the above, the EEE (at zero EEE_0 and at infinity EEE_∞) becomes

$$\text{EEE}_0 = \lim_{\rho\to 0} \text{EEE} = 0.$$

$$\text{EEE}_\infty = \lim_{\rho\to\infty} \text{EEE} = 0.$$

which concludes the proof.

REFERENCES

[1] Ericsson, "5G Wireless Access: An Overview," *Ericsson White Paper*, April 2020.

[2] J. A. Ansere *et al.*, "Optimal Resource Allocation in Energy-Efficient Internet-of-Things Networks With Imperfect CSI," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5401–5411, 2020.

[3] H. Seo, J. Hong, and W. Choi, "Low Latency Random Access for Sporadic MTC Devices in Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5108–5118, 2019.

[4] L. Chettri and R. Bera, "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 16–32, 2020.

[5] A. H. Bui *et al.*, "A Comprehensive Distributed Queue-Based Random Access Framework for mMTC in LTE/LTE-A Networks With Mixed-Type Traffic," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 107–12 120, 2019.

[6] E. Soltanmohammadi et al., "A Survey of Traffic Issues in Machine-to-Machine Communications Over LTE," *IEEE Internet of Things Journal*, Dec 2016.

[7] J. Navarro-Ortiz *et al.*, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[8] B. K. J. Al-Shammari, N. Al-Aboody, and H. S. Al-Raweshidy, "IoT Traffic Management and Integration in the QoS Supported Network," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 352–370, 2018.

[9] 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.

[10] G. Durisi, T. Koch, and P. Popovski, "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication with Short Packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[11] W. Yang *et al.*, "Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, July 2014.

[12] G. Ozcan and M. C. Gursoy, "Throughput of Cognitive Radio Systems with Finite Blocklength Codes," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2541–2554, November 2013.

[13] M. Haghifam *et al.*, "Joint Sum Rate and Error Probability Optimization: Finite Blocklength Analysis," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 726–729, 2017.

[14] I. Gravalos *et al.*, "Efficient Network Planning for Internet of Things With QoS Constraints," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3823–3836, 2018.

[15] W. Cheng, X. Zhang, and H. Zhang, "Heterogeneous Statistical QoS Provisioning for Downlink Transmissions over Mobile Wireless Cellular Networks," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 4622–4628.

[16] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective Capacity in Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3007–3038, 2019.

[17] B. Makki *et al.*, "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.

[18] F. Qasmi *et al.*, "Optimum Transmission Rate in Fading Channels with Markovian Sources and QoS Constraints," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–5.

[19] M. C. Gursoy, "Throughput Analysis of Buffer-Constrained Wireless Systems in the Finite Blocklength Regime," in *2011 IEEE International Conference on Communications (ICC)*, June 2011, pp. 1–5.

[20] Z. Dong *et al.*, "Energy Efficiency Optimization and Resource Allocation of Cross-Layer Broadband Wireless Communication System," *IEEE Access*, vol. 8, pp. 50 740–50 754, 2020.

[21] M. Shehab *et al.*, "Effective Energy Efficiency of Ultra-reliable Low Latency Communication," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[22] Q. Ye *et al.*, "Traffic-Load-Adaptive Medium Access Control for Fully Connected Mobile Ad Hoc Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9358–9371, 2016.

[23] 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, 2015.

[24] B. Farayev, Y. Sadi, and S. C. Ergen, "Optimal Power Control and Rate Adaptation for Ultra-Reliable M2M Control Applications," in *2015 IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6.

[25] 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.

[26] O. L. Alcaraz Lopez, H. Alves, and M. Latva-Aho, "Rate Control under Finite Blocklength for Downlink Cellular Networks with Reliability Constraints," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, 2018, pp. 1–6.

[27] M. Sinaie *et al.*, "A Novel Power Consumption Model for Effective Energy Efficiency in Wireless Networks," *IEEE Wireless Communications Letters*, vol. 5, no. 2, pp. 152–155, April 2016.

[28] J. Xu *et al.*, "Use of Two-Mode Transceiver Circuitry and Its Cross-Layer Energy Efficiency Analysis," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2065–2068, 2017.

[29] S. Schiessl *et al.*, "Delay Performance of Wireless Communications With Imperfect CSI and Finite-Length Coding," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6527–6541, 2018.

[30] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. Dover, 1965.

[31] M. Ozmen and M. C. Gursoy, "Wireless Throughput and Energy Efficiency with Random Arrivals and Statistical Queuing Constraints," *IEEE Transactions on Information Theory*, March 2016.

[32] Dapeng Wu and R. Negi, "Effective Capacity: a Wireless Link Model for Support of Quality of Service," *IEEE Transactions on Wireless Communications*, July 2003.

[33] B. Makki, T. Svensson, and M. Zorzi, "Finite Block-Length Analysis of the Incremental Redundancy HARQ," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, 2014.

[34] N. H. Mahmood *et al.*, "Six Key Features of Machine Type Communication in 6G," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[35] O. L. A. López *et al.*, "Ultra-Low Latency, Low Energy, and Massiveness in the 6G Era via Efficient CSIT-Limited Scheme," *IEEE Communications Magazine*, vol. 58, no. 11, pp. 56–61, 2020.

[36] M. C. Gursoy, "Throughput Analysis of Buffer-Constrained Wireless Systems in the Finite Blocklength Regime," *EURASIP Journal on Wireless Communications and Networking*, 2013.

[37] J. Wu *et al.*, "Base-Station Sleeping Control and Power Matching for Energy-Delay Tradeoffs With Bursty Traffic," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3657–3675, 2016.

[38] L. Sboui *et al.*, "A New Relation Between Energy Efficiency and Spectral Efficiency in Wireless Communications Systems," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 168–174, 2019.

[39] K. Singh, M.-L. Ku, and M. F. Flanagan, "Energy-Efficient Precoder Design for URLLC-Enabled Downlink Multi-User MISO Networks Using Finite Blocklength Codes," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[40] B. Makki *et al.*, "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.

[41] 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.

[42] S. Popli, R. K. Jha, and S. Jain, "A Survey on Energy Efficient Narrowband Internet of Things (NBIoT): Architecture, Application and Challenges," *IEEE Access*, vol. 7, pp. 16 739–16 776, 2019.

[43] J. M. d. S. SantAna *et al.*, "Hybrid Coded Replication in LoRa Networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5577–5585, 2020.

[44] A. Pizzo, A. Zappone, and L. Sanguinetti, "Solving Fractional Polynomial Problems by Polynomial Optimization Theory," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1540–1544, 2018.

[45] L. Shi *et al.*, "Computation Energy Efficiency Maximization for a NOMA Based WPT-MEC Network," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[46] M. Guo and M. C. Gursoy, "Energy-Efficient Joint Antenna and User Selection in Single-Cell Massive MIMO System," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 838–842.

[47] C. Isheden *et al.*, "Framework for Link-Level Energy Efficiency Optimization with Informed Transmitter," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2946–2957, 2012.

[48] X. Gao *et al.*, "Machine Learning Inspired Energy-Efficient Hybrid Precoding for mmWave Massive MIMO Systems," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[49] J. Huang *et al.*, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 288–296, 2009.

[50] S. D'Oro *et al.*, "A learning approach for low-complexity optimization of energy efficiency in multicarrier wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3226–3241, 2018.

[51] Z. Wu and M. Kolonko, "Asymptotic Properties of a Generalized Cross-Entropy Optimization Algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 5, pp. 658–673, 2014.

[52] F. Dedov, *The Bible of Algorithms and Data Structures: A Complex Subject Simply Explained (Runtime Complexity, Big O Notation, Programming)*. Amazon Digital Services LLC - KDP Print US, 2020. [Online]. Available: https://books.google.fi/books?id=FB_BzQEACAAJ