Fu *et al. J Wireless Com Network* (2021) 2021:26
https://doi.org/10.1186/s13638-021-01905-7

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

# Energy-efficient offloading and resource allocation for mobile edge computing enabled mission-critical internet-of-things systems

Yaru Fu[1], Xiaolong Yang[3*], Peng Yang[2], Angus K. Y. Wong[1], Zheng Shi[4], Hong Wang[5,6] and Tony Q. S. Quek[2]

*Correspondence: xiaolongyang@bistu.edu.cn
[3] School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China
Full list of author information is available at the end of the article

**Abstract**

The energy cost minimization for mission-critical internet-of-things (IoT) in mobile edge computing (MEC) system is investigated in this work. Therein, short data packets are transmitted between the IoT devices and the access points (APs) to reduce transmission latency and prolong the battery life of the IoT devices. The effects of short-packet transmission on the radio resource allocation is explicitly revealed. We mathematically formulate the energy cost minimization problem as a mixed-integer non-linear programming (MINLP) problem, which is difficult to solve in an optimal way. More specifically, the difficulty is essentially derived from the coupling of the binary offloading variables and the resource management among all the IoT devices. For analytical tractability, we decouple the mixed-integer and non-convex optimization problem into two sub-problems, namely, the task offloading decision-making and the resource optimization problems, respectively. It is proved that the resource allocation problem for IoT devices under the fixed offloading strategy is convex. On this basis, an iterative algorithm is designed, whose performance is comparable to the best solution for exhaustive search, and aims to jointly optimize the offloading strategy and resource allocation. Simulation results verify the convergence performance and energy-saving function of the designed joint optimization algorithm. Compared with the extensive baselines under comprehensive parameter settings, the algorithm has better energy-saving effects.

**Keywords:** Energy minimization, Internet-of-things (IoTs), Mobile edge computing (MEC), Offloading decision, Resource management, Short packet transmission

## 1 Introduction

The rapid growth of innovation applications such as smart agriculture, smart factories, and intelligent traffic monitoring system has triggered an explosive growth of internet-of-things (IoT) devices [1, 2]. Nevertheless, IoT devices are usually limited by batteries and computing power, or even no computing power, which makes IoT devices unable to process data. To solve these problems, mobile edge computing (MEC) has been proposed as a promising solution [3–6]. The main idea of MEC is to complete the computation-intensive and energy-intensive tasks of IoT devices at the edge nodes, such as access points (APs) or base stations, through task offloading. After the MEC node completes

the data processing, the calculation results (e.g., working or operation instructions to the IoT terminals) will be returned to the devices. In this regard, the energy consumption of the devices can be significantly reduced, which can reduce the device maintenance costs in actual IoT usage scenarios. MEC enabled mission-critical IoT system has been taken as a key enabler for the sustainability of wireless cellular networks.

Extensive work has been done to study the joint task offloading and radio resource optimization for MEC-enabled IoT systems [7–9]. Specifically, in [7], the total summation of the hovering energy and the computation energy minimization problem for the MEC system assisted by unmanned aerial vehicle (UAV) was studied. Therein, the hovering time, scheduling and resource allocation for IoT users are jointly optimized. In addition, the authors in [8] solved the problem of offloading decision for multiple users from the perspective of game theory. By designing the optimal response algorithm, it is proved that there is a Nash equilibrium among users. In addition, in the literature [9], the offloading of the MEC network supporting Orthogonal Frequency Division Multiple Access (OFDMA) and the joint uplink and downlink resource allocation problems were studied. Therein, several sub-optimal solutions that achieve different leverages between the system performance and the time complexity are explored. Recently, machine learning aware methods have been widely utilized to make task offloading decisions on MEC networks [10–12]. More precisely, Yang et al. studied the joint communication and computation resource allocation for non-orthogonal multiple access (NOMA) assisted MEC systems in [10]. A multi-agent Q-learning empowered algorithm is developed to make offloading decisions for multiple NOMA users. Moreover, Huang et al. in [11] explored a deep reinforcement learning algorithm to perform the online computation offloading for wireless powered MEC systems. It was explicitly shown that the designed learning assisted scheme makes the real-time and optimal offloading become a reality. Furthermore, Wu et al. [12] investigated the combination between the optimization method and the Q-learning strategy to optimize the joint offloading and resource allocation problem in MEC systems, while considering the time-varying channel and unknown users' channel state information.

However, the aforementioned work [7–12] mainly focused on the conventional packet transmission scheme, in which the Shannon capacity formula was applied. Nevertheless, for mission-critical IoT, enabling the transmission mechanism of the conventional block length mechanism may cause a longer delay, which further shortens the battery life of IoT devices [13]. To meet the stringent requirements of latency-sensitive applications, a MEC network assisted by short packet transmission is proposed. For which, several remarkable attempts are worth to be introduced [14–16]. More specifically, the authors in [14] studied the offloading decision for MEC systems supporting single-carrier time division multiple access. Meanwhile, in [15, 16], the resource allocation for MEC systems supporting OFDMA was analyzed with short packet transmission consideration. In particular, a deep learning based approach was utilized in [15] to allocate online resources for MEC users to minimize the maximum normalized energy consumption of each user. In addition, the optimal resource management of OFDMA-MEC system for ultra-reliable low-latency communication (URLLC) was studied in [16]. The goal in [16] was to minimize the end-to-end latency of the network. Thereof, the optimal joint optimization solution based on branch and bound method and two time-efficient suboptimal

solutions are designed, respectively. Furthermore, the interplay between URLLC and enhanced mobile broadband (eMBB) has been studied by Yang et al. in [17], where the energy efficiency and the sum of received SNRs are maximized for eMBB users and URLLC users, respectively. The formulated multiple timescales problem has been solved by a sample average approximation (SAA) oriented technique. Similar models to that in [17] have been investigated by Tang et al. [18] and Anand et al. [19] to minimize the URLLC users' latency.

Since the next generation of wireless cellular networks has strict requirements on the energy efficiency and transmission efficiency (low latency) [20–23], in this paper, we investigate the short-packet transmission enabled data offloading between IoT devices and access points, aiming at minimizing the total energy consumption. Our main contributions are summarized as follows:

- We formulated the energy cost minimization problem for the short packets transmission for mission-critical IoT in MEC system, and revealed the effect of short packet transmission on radio resource management. The minimization problem is a mixed integer non-linear programming (MINLP) problem, and it is challenging to solve this problem in the best way. The difficulty is essentially derived from the coupling of offloading strategy and resource optimization.
- For ease of tractability, we decoupled the original MINLP problems into two subproblems: 1) the offloading decision-making sub-problem and 2) the resource allocation sub-problem. It clearly demonstrated that the resource allocation sub-problem of IoT devices under the fixed offloading policy is a convex optimization problem, and its optimal solution can be realized by various convex optimization tools.
- Through the previous analysis, a time-saving algorithm that works in an iterative manner is developed to jointly optimize the offloading strategy and the resource (i.e., computation, power and bandwidth), so as to have a provable convergence guarantee.

Apart from the above contributions, Monte-Carlo simulations were also conducted to verify the effiectiveness of our proposed joint optimization scheme from various aspects, such as the convergence and the energy consumption, under extensive and comprehensive system settings.

The rest of this paper is summarized as follows: in Sect. 2, we introduce the considered MEC-enabled mission-critical IoT system model and discuss the total energy consumption of the system. In Sect. 3, the problem of minimizing energy cost is formulated, which considers the joint optimization of the offloading decision, bandwidth, and transmit power. In Sect. 4, we propose a time-efficient near optimal algorithm to solve the non-tractable MINLP problem. Monte-Carlo simulation is performed in Sect. 5 to evaluate the performance of our developed strategy in terms of convergence performance and total energy consumption. Finally, we conclude this paper and reveal the future research direction.

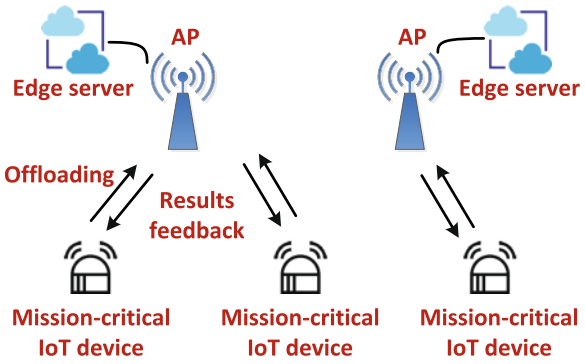


**Fig. 1** MEC-enabled mission-critical IoT system model

## 2 System description and energy consumption

In this section, we first introduce the system model of the mobile edge computing (MEC) enabled mission-critical internet of things (IoT) network. Whereafter, we elaborate on the achievable data rate of each IoT device and the system's energy cost.

### 2.1 System model

We consider the MEC based IoT network that consists of $N$ APs serving $K$ devices, as illustrated by Fig. 1. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$ be the index sets of all the APs and the IoT devices, respectively. Each of the IoT devices has a latency-critical task to process. Denote by $T_k$ the delay requirement of the $k$th IoT terminal, where $k \in \mathcal{K}$. The APs provide computing services to all the devices as the IoT devices have no data processing capability in the context of this article.

In addition, it is assumed that both the APs and the devices are equipped with a single antenna. For $k \in \mathcal{K}$, we use a three-dimension tuple to characterize the task of the $k$th device, referred to as $(\mathbb{U}_k, \mathbb{F}_k, \mathbb{D}_k)$. Specifically, $\mathbb{U}_k$ indicates the data volume of device $k$ that needs to be offloaded/computed. In addition, $\mathbb{F}_k$ (in CPU cycles) denotes the computation resource required by the $k$th device to execute its data. Moreover, $\mathbb{D}_k$ represents the output data size of device $k$ after the processing is completed at its selected AP.

We stipulate that each device chooses one of the APs to offload its task for further process.[1] To characterize this feature, we define an auxiliary variable, referred to as $a_{k,n} \in \{0, 1\}$, as the indicator to show whether the $k$th device uploads its data to AP $n$ or not, where $k \in \mathcal{K}$ and $n \in \mathcal{N}$. In detail, $a_{k,n} = 1$ if device $k$ is associated with the $n$th AP and $a_{k,n} = 0$ otherwise. With aforementioned definitions, we have the following constraint,

$$\sum_{n \in \mathcal{N}} a_{k,n} = 1, \tag{1}$$

where $k \in \mathcal{K}$. Before ending this subsection, we define $\tau_{\mathrm{U}}$ and $\tau_{\mathrm{D}}$ as the offloading transmission time and the downloading transmission time of IoT devices, respectively.

[1] In this paper, the security and privacy issues are not considered. Namely, it is stipulated that all the processes of data sharing are in the case of information security.

### 2.2 Achievable capacity

As mentioned, we focus on the short packet transmission between the IoT devices and the APs. The upload and download transmission rates of the $k$th device are denoted by $R^{\mathrm{U}}_{k,n}$ and $R^{\mathrm{D}}_{k,n}$, respectively, therein we assume that device $k$ is associated with AP $n$. In addition, for $k \in \mathcal{K}$ and $n \in \mathcal{N}$, define $g_{k,n}$ as the channel gain between IoT device $k$ and AP $n$. Based on [24, 25], the maximum achievable rate in short packet regime can be accurately approximated by

$$R^{\mathrm{U}}_{k,n} \approx \frac{W^{\mathrm{U}}_{k,n}}{\ln 2}\left[\ln(1+\frac{g_{k,n}p^{\mathrm{U}}_{k,n}}{\delta N_0 W^{\mathrm{U}}_{k,n}}) - \sqrt{\frac{V^{\mathrm{U}}_{k,n}}{\tau_{\mathrm{U}} W^{\mathrm{U}}_{k,n}}}Q_G^{-1}(\varepsilon_k)\right], \tag{2}$$

where $W^{\mathrm{U}}_{k,n}$ and $p^{\mathrm{U}}_{k,n}$ indicate the assigned bandwidth and transmit power to device $k$ for offloading data to AP $n$, respectively. $\delta$ represents the signal-to-noise-ratio (SNR) loss due to imperfect channel state information (CSI) at the transmitter [26], $N_0$ depicts the single-side noise power spectrum density. Moreover, $\varepsilon_k$ is the decoding error probability of device $k$. Furthermore, $Q_G^{-1}(\cdot)$ expresses the inverse of Gaussian Q-function and $V^{\mathrm{U}}_{k,n}$ is the channel dispersion of device $k$ [27], which is quoted as follows:

$$V^{\mathrm{U}}_{k,n} = 1 - \frac{1}{(1+\frac{g_{k,n}p^{\mathrm{U}}_{k,n}}{\delta N_0 W^{\mathrm{U}}_{k,n}})^2}. \tag{3}$$

Note that $V^{\mathrm{U}}_{k,n}$ is approximated to 1 in the following. Such an approximation is very accuracy, when the received SNR is higher than 5 dB, which has been precisely validated by the authors in [25]. Details are omitted here to avoid redundancy.

### 2.3 Energy consumption analysis

In this subsection, we elaborate on the energy consumption induced by the IoT devices' data offloading and the computation results downloading, as well as the data processing at the APs, respectively.

#### *2.3.1 Energy consumption of offloading and downloading*

We assume that device $k$ selects AP $n$ to do the data offloading. With aforementioned definitions, we have

$$R^{\mathrm{U}}_{k,n}\tau_{\mathrm{U}} \geq \mathbb{U}_k. \tag{4}$$

Substituting (2) into (4), the transmit power of device $k$, i.e., $p^{\mathrm{U}}_{k,n}$, should satisfy

$$\begin{aligned} p^{\mathrm{U}}_{k,n} &\geq \frac{\delta N_0 W^{\mathrm{U}}_{k,n}}{g_{k,n}}\left\{\exp(\frac{\mathbb{U}_k \ln 2}{\tau_U W^{\mathrm{U}}_{k,n}} + \frac{Q_G^{-1}(\varepsilon_k)}{\sqrt{\tau_{\mathrm{U}} W^{\mathrm{U}}_{k,n}}}) - 1\right\} \\ &= f^{\mathrm{U}}_k(W^{\mathrm{U}}_{k,n}). \end{aligned} \tag{5}$$

For simplicity, we use $f_k^{\mathrm{U}}(W_{k,n}^{\mathrm{U}})$ to represent the lower bound of $p_{k,n}^{\mathrm{U}}$. It is noteworthy that $f_k^{\mathrm{U}}(W_{k,n}^{\mathrm{U}})$ is non-convex with respect to (w.r.t.) $W_{k,n}^{\mathrm{U}}$ [25]. In a similar way, the constraint for the downlink phase can be expressed as follows:

$$R_{k,n}^{\mathrm{D}}\tau_{\mathrm{D}} \geq \mathbb{D}_k, \tag{6}$$

from which the allocated transmission power to device $k$ in the download phase can be obtained, namely,

$$\begin{aligned} p_{k,n}^{\mathrm{D}} &\geq \frac{\delta N_0 W_{k,n}^{\mathrm{D}}}{g_{k,n}} \left\{ \exp\left(\frac{\mathbb{D}_k \ln 2}{\tau_{\mathrm{D}} W_{k,n}^{\mathrm{D}}} + \frac{Q_G^{-1}(\varepsilon_k)}{\sqrt{\tau_{\mathrm{D}} W_{k,n}^{\mathrm{D}}}}\right) - 1 \right\} \\ &= f_k^{\mathrm{D}}(W_{k,n}^{\mathrm{D}}), \end{aligned} \tag{7}$$

where $W_{k,n}^{\mathrm{D}}$ and $p_{k,n}^{\mathrm{D}}$ express the allocated bandwidth and transmit power for device $k$ to download the computation results from AP $n$, respectively. It is not hard to check that $f_k^{\mathrm{D}}(W_{k,n}^{\mathrm{D}})$ is also a non-convex function w.r.t. $W_{k,n}^{\mathrm{D}}$.

Although $f_k^{\mathrm{U}}(W_{k,n}^{\mathrm{U}})$ and $f_k^{\mathrm{D}}(W_{k,n}^{\mathrm{D}})$ are non-convex w.r.t. $W_{k,n}^{\mathrm{U}}$ and $W_{k,n}^{\mathrm{D}}$, they have some interesting properties as quoted below [13]:

**Property 1** *There is a unique solution* $\tilde{W}_{k,n}^{U}$ $(\tilde{W}_{k,n}^{D})$ *that minimizes* $f_k^{U}(W_{k,n}^{U})$ $(f_k^{D}(W_{k,n}^{D}))$. *In addition,* $f_k^{U}(W_{k,n}^{U})$ $(f_k^{D}(W_{k,n}^{D}))$ *is strictly convex w.r.t.* $W_{k,n}^{U}$ $(W_{k,n}^{D})$ *when* $0 < W_{k,n}^{U} \leq \tilde{W}_{k,n}^{U}$ $(0 < W_{k,n}^{D} \leq \tilde{W}_{k,n}^{D})$.

Based on the foregoing property, we have the following Lemma:

**Lemma 1** *The global optimal solution of problem:*

$$\min_{W_{k,n}^{U}, W_{k,n}^{D}, p_{k,n}^{U}, p_{k,n}^{D}} \mathscr{F}\left(W_{k,n}^{U}, W_{k,n}^{D}, p_{k,n}^{U}, p_{k,n}^{D}\right)$$

*with constraints* (3), (4) and $W_{k,n}^{m} > 0$, $p_{k,n}^{m} > 0$, $m \in \{U, D\}$, *can be obtained via solving the optimization problem with objective function*

$$\min_{W_{k,n}^{\mathrm{U}}, W_{k,n}^{\mathrm{D}}, p_{k,n}^{U}, p_{k,n}^{D}} \mathscr{F}\left(W_{k,n}^{U}, W_{k,n}^{D}, p_{k,n}^{U}, p_{k,n}^{D}\right),$$

*subjecting to*

$$(3),\ (4),\ W_{k,n}^{m} > 0,\ p_{k,n}^{m} > 0,\ W_{k,n}^{m} \leq \tilde{W}_{k,n}^{m},\ m \in \{U, D\},$$

*given that the objective function* $\mathscr{F}$ *is increases with bandwidth* $W_{k,n}^{m}$ *and transmit power* $p_{k,n}^{m}$, *where* $m \in \{U, D\}$.

The aforementioned conclusions have been proved by [13]. Details are omitted here for brevity. Denote by $E_{k,n}^{\mathrm{U}}$ the energy cost of device $k$ in terms of offloading data to the $n$th AP. In addition, let $E_{k,n}^{\mathrm{D}}$ be the energy consumption of the $n$th AP to feedback the computation results to device $k$. As a consequence, we have

$$E_{k,n}^{\mathrm{U}} = a_{k,n} p_{k,n}^{\mathrm{U}} \tau_{\mathrm{U}},\ k \in \mathcal{K}, \tag{8}$$

and

$$E_{k,n}^{\mathrm{D}} = a_{k,n} p_{k,n}^{\mathrm{D}} \tau_{\mathrm{D}},\ k \in \mathcal{K}. \tag{9}$$

#### *2.3.2 Energy cost of APs for data processing*

Define $L_{k,n}$ as the assigned computation resource for IoT device $k$ by AP $n$. Thereby, the execution delay of device $k$, denoted by $\tau_{k,n}$, is obtained as follows:

$$\tau_{k,n} = \mathbb{F}_k / L_{k,n},\ k \in \mathcal{K}. \tag{10}$$

It is assumed that the AP executes its associated devices' tasks in a sequential manner.[2] Thereby, the following latency requirement should be met:

$$a_{k,n}(\tau_{\mathrm{U}} + \tau_{\mathrm{D}} + \sum_{k \in \mathcal{K}} \tau_{k,n}) \leq T_k,\ n \in \mathcal{N},\ k \in \mathcal{K}. \tag{11}$$

In addition, define $L_n$ (in cycles) as the computation capacity of AP $n$, where $n \in \mathcal{N}$. With aforementioned definitions, we have

$$\sum_{k \in \mathcal{K}} a_{k,n} L_{k,n} \leq L_n,\ n \in \mathcal{N}. \tag{12}$$

The energy consumed by data processing for device $k$ at AP $n$, denoted by $E_{k,n}^{\mathrm{P}}$, can be expressed as follows:

$$E_{k,n}^{\mathrm{P}} = a_{k,n} \alpha_n \mathbb{F}_k (L_{k,n})^2, \tag{13}$$

where $\alpha_n$ indicates the effective switched capacitance, which is a hardware architecture dependent parameter [7]. Following the system setup in [7], we let $\alpha_n = 10^{-27}$ for $n \in \mathcal{N}$ .

Define $E_{\mathrm{Total}}$ as the total energy expenditure of the whole system. With the foregoing analysis, we have

$$E_{\mathrm{Total}} = \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} (E_{k,n}^{\mathrm{U}} + E_{k,n}^{\mathrm{D}} + E_{k,n}^{\mathrm{P}}), \tag{14}$$

where $E_{k,n}^{\mathrm{U}}$, $E_{k,n}^{\mathrm{D}}$, and $E_{k,n}^{\mathrm{P}}$ have been given in (8), (9), and (13), respectively.

## 3 Problem formulation

In this section, we discuss the considered optimization problem formulation. Some definitions are given at first. Denote by $\boldsymbol{W} = (W_{k,n}^{\mathrm{U}}, W_{k,n}^{\mathrm{D}})_{k \in \mathcal{K}, n \in \mathcal{N}}$ the bandwidth allocation for all the devices. In addition, define $\boldsymbol{p} = (p_{k,n}^{\mathrm{U}}, p_{k,n}^{\mathrm{D}})_{k \in \mathcal{K}, n \in \mathcal{N}}$ as the power allocation for all the IoT devices. Besides, denote $\boldsymbol{L} = (L_{k,n})_{k \in \mathcal{K}, n \in \mathcal{N}}$ as the computation resource allocation strategy of the system. Moreover, let $\boldsymbol{a}_k = (a_{k,n})_{n \in \mathcal{N}}$ and $\boldsymbol{a} = (\boldsymbol{a}_k)_{k \in \mathcal{K}}$ be the

[2] It is worth mentioning that we follow this data processing model of APs by [8, 9].

offloading decision vector of device $k$ and the system, respectively. Furthermore, denote by $W$ the total system bandwidth, while let $P_n$ be the transmit power budget of AP $n$, where $n \in \mathcal{N}$. In this work, we aim to minimize the total energy cost via jointly optimizing the offloading decision, computation resource, bandwidth, as well as the transmit power, taking into account the total computation capability, bandwidth and power constraints. The minimization problem, referred to as $\mathbf{P}(0)$, is mathematically formulated as follows:

$$\mathbf{P}(0): \min_{\boldsymbol{W},\boldsymbol{p},\boldsymbol{L},\boldsymbol{a}} E_{\text{Total}} \tag{15}$$

s.t.

$$\begin{aligned}
&\text{C1}: p_{k,n}^{\text{U}} \geq a_{k,n} f_k^{\text{U}}(W_{k,n}^{\text{U}}),\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C2}: p_{k,n}^{\text{D}} \geq a_{k,n} f_k^{\text{D}}(W_{k,n}^{\text{D}}),\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C3}: a_{k,n}(\tau_{\text{U}} + \tau_{\text{D}} + \sum_{k \in \mathcal{K}} \tau_{k,n}) \leq T_k,\ n \in \mathcal{N},\ k \in \mathcal{K},\\
&\text{C4}: \sum_{k \in \mathcal{K}} a_{k,n} W_{k,n}^{\text{U}} \leq W,\ n \in \mathcal{N}\\
&\text{C5}: \sum_{k \in \mathcal{K}} a_{k,n} W_{k,n}^{\text{D}} \leq W,\ n \in \mathcal{N}\\
&\text{C6}: \sum_{k \in \mathcal{K}} a_{k,n} p_{k,n}^{\text{D}} \leq P_n,\ n \in \mathcal{N},\\
&\text{C7}: \sum_{k \in \mathcal{K}} a_{k,n} L_{k,n} \leq L_n,\ n \in \mathcal{N},\\
&\text{C8}: p_{k,n}^{\text{U}} \geq 0,\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C9}: p_{k,n}^{\text{D}} \geq 0,\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C10}: W_{k,n}^{\text{U}} \geq 0,\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C11}: W_{k,n}^{\text{D}} \geq 0,\ k \in \mathcal{K},\ n \in \mathcal{N},\\
&\text{C12}: a_{k,n} \in \{0,1\},\ k \in \mathcal{K}, n \in \mathcal{N},\\
&\text{C13}: \sum_{n \in \mathcal{N}} a_{k,n} = 1,\ k \in \mathcal{K},
\end{aligned}$$

where C1 and C2 indicate the power constraints of the offloading and the downloading phases, respectively. C3 represents the latency requirement of each IoT device. C4 as well as C5, and C6 express the total bandwidth requirement and the power budget of each AP, respectively. C7 gives the computation resource constraint of each AP. C8–C9 and C10–C11 depict the non-negativity of the transmit power and the allocated bandwidth, respectively. C12 illustrates the binarity of the offloading decision. Moreover, C13 indicates that each device can only select one AP to do the task offloading. The energy minimization problem $\mathbf{P}(0)$ is a mixed-integer non-linear programming (MINLP) problem, which is in general NP-hard and is difficult to solve [7]. For ease of tractability, in the following section, a time-efficient sub-optimal algorithm with ensured convergence performance is proposed to do the joint offloading and resource management for IoT devices.

## 4 Methodology of resource management

In this section, we first demonstrate that the resultant resource allocation problem under given offloading decision is convex. With which, we show that the optimal solution of $\mathbf{P}(0)$ can be obtained via an exhaustive search oriented method, which suffers from a worst-case exponentially increased time complexity. Subsequently, a time-efficient joint offloading decision and resource allocation algorithm is proposed, which works in an iterative manner and has ensured convergence performance.

### 4.1 Resource allocation with fixed offloading decision

Given the offloading strategy of each IoT device, the original energy minimization problem $\mathbf{P}(0)$, can be re-written as a resource management problem, denoted by $\mathbf{P}(1)$, and it is given as follows:

$$\mathbf{P}(1): \min_{\boldsymbol{W},\boldsymbol{p},\boldsymbol{L}} E_{\text{Total}} \tag{16}$$

subject to C1 to C11.

**Lemma 2** $\mathbf{P}(1)$ *is a convex optimization problem.*

***Proof*** To show that problem $\mathbf{P}(1)$ is convex, we only need to clarify the convexity of the objective function as well as the constraints C1 and C2 since the convexity of the other constraints is obvious.

According to (8), (9), and (13), it is easy to check that $E^{\text{U}}_{k,n}$, $E^{\text{D}}_{k,n}$ and $E^{\text{P}}_{k,n}$ are all convex w.r.t. the variables $(\boldsymbol{W},\boldsymbol{p},\boldsymbol{L})$. Namely, the objective function is convex. In addition, it increases with $\boldsymbol{W}$ and $\boldsymbol{p}$, which means C1 and C2 are equivalent to the constraints C1, C2 and $W^m_{k,n} \leq \tilde{W}^m_{k,n}$, where $m \in \{\text{U},\text{D}\}$, according to Lemma 2. Together with Property 1, the proof is completed. □

With aforementioned analysis, $\mathbf{P}(0)$ can be optimally solved by various existing convex optimization tools. In other words, the optimal joint offloading decision and resource allocation method for problem $\mathbf{P}(0)$ can be obtained via a full search methodology.

**Lemma 3** *The optimal solution for* $\mathbf{P}(0)$ *suffers an exponentially increased time complexity.*

***Proof*** The computation complexity of the full-search enabled optimal solution relies on the size of the strategy space, which is exponentially increased by $K$. The proof is completed. □

Since the exhaustive search scheme suffers from an exponentially increased computation complexity, which is not affordable for practical IoT systems, especially when the number of the devices is large. In the following subsection, a time-efficient algorithm is developed, which is implemented in an iterative manner and has ensured convergence property.

### 4.2 Time-efficient joint offloading strategy and resource management algorithm

In this subsection, we investigate a sub-optimal joint offloading decision and resource allocation method. Define $\boldsymbol{e}_n$ as a $N$-dimension vector, whose items are all zeros except the $n$th component, which is set to be 1. In addition, denote by $\mathcal{E} = \{\boldsymbol{e}_n | n \in \mathcal{N}\}$. Obviously, we have $\boldsymbol{a}_k \subset \mathcal{E}$. Moreover, define $\boldsymbol{a}_k^t$ and $\boldsymbol{a}^t = (\boldsymbol{a}_k^t)_{k\in\mathcal{K}}$ as the offloading decision of IoT device $k$ and the offloading strategy of the system in the $t$th iteration, respectively. The joint optimization algorithm is conducted in an iterative manner. In the $t$th iteration, we first determine the optimal offloading scheme for device $k$, denoted by $\bar{\boldsymbol{a}}_k^t$ and it is given below

$$\bar{\boldsymbol{a}}_k^t = \underset{\boldsymbol{a}_k^t \subset \mathcal{E}}{\operatorname{argmin}} E_{\text{Total}}(\boldsymbol{a}_{-k}^t), \tag{17}$$

where $E_{\text{Total}}$ represents the system energy consumption under the optimal resource mangement.[3] In addition, $\boldsymbol{a}_{-k}^t$ is defined as follows:

$$\boldsymbol{a}_{-k}^t = (\boldsymbol{a}_1^{t-1}, \boldsymbol{a}_2^{t-1}, \ldots, \boldsymbol{a}_{k-1}^{t-1}, \boldsymbol{a}_k^t, \boldsymbol{a}_{k+1}^{t-1}, \ldots, \boldsymbol{a}_K^{t-1}).$$

Moreover, we denote by

$$\bar{\boldsymbol{a}}_{-k}^t = (\boldsymbol{a}_1^{t-1}, \boldsymbol{a}_2^{t-1}, \ldots, \boldsymbol{a}_{k-1}^{t-1}, \bar{\boldsymbol{a}}_k^t, \boldsymbol{a}_{k+1}^{t-1}, \ldots, \boldsymbol{a}_K^{t-1}). \tag{18}$$

Furthermore, let $k^*$ be the user that satisfies

$$k^* \triangleq \underset{k\in\mathcal{K}}{\operatorname{argmin}} E_{\text{Total}}(\bar{\boldsymbol{a}}_{-k}^t). \tag{19}$$

With aforementioned definitions, we renew the offloading decision of the system in $t$th iteration as follows:

$$\boldsymbol{a}^t = \bar{\boldsymbol{a}}_{-k^*}^t, \tag{20}$$

until the stopping criteria are satisfied. For brevity, we summarize the pseudo-code of our sub-optimal joint offloading and resource allocation scheme in Algorithm 1.

[3] Given the offloading decision strategy of the system, the energy minimization problem becomes convex and can be solve via any kind of convex optimization tools. In this work, the MATLAB CVX tool is applied.

**Algorithm 1** The suboptimal algorithm for $\mathbf{P}(0)$

1: Given the initial offloading decision vector $\boldsymbol{a}^0$. Let $t = 1$. In addition, set the maximum iteration number as $\bar{T}$.
2: **repeat**
3: For $k \in \mathcal{K}$, determine the best strategy for device $k$. Namely,

$$\bar{\boldsymbol{a}}_k^t = \underset{\boldsymbol{a}_k^t \subset \mathcal{E}}{\operatorname{argmin}} E_{\text{Total}}(\boldsymbol{a}_{-k}^t).$$

4: Select user $k^*$ that induces to the minimum system energy consumption, i.e.,

$$k^* \triangleq \underset{k \in \mathcal{K}}{\operatorname{argmin}} E_{\text{Total}}(\bar{\boldsymbol{a}}_{-k}^t).$$

5: Update the offloading strategy for iteration $t$ as follows:

$$\boldsymbol{a}^t = \bar{\boldsymbol{a}}_{-k^*}^t.$$

6: **until** $E(\boldsymbol{a}^t) > E(\boldsymbol{a}^{t-1})$ or $t > \bar{T}$.
7: **return** the offloading strategy of the system, i.e., $\boldsymbol{a}^{(t)}$ and the optimal resource allocation strategy of problem $\mathbf{P}(0)$ with $\boldsymbol{a} = \boldsymbol{a}^{(t)}$.

The convergence performance of our designed joint optimization approach is discussed below:

**Lemma 4** *The convergence of the objective function in our developed joint offloading and resource management algorithm is ensured.*

***Proof*** The proof for the convergence performance is obvious since the system's energy consumption is degraded with the iterations. In addition, the total energy cost is lower bounded by zero. The proof is completed. □

Before ending this section, we analyze the computational complexity of our developed algorithm. Details are expressed in the following lemma:

**Lemma 5** *The time complexity of Algorithm 1 is* $\mathcal{O}(KN)$.

***Proof*** Based on Algorithm 1, to update the system's offloading strategy, we need to calculate the energy consumption under all the possible offloading policies for each IoT device, i.e., (17). The maximum number of the offloading schemes per device is $N$. In total, we have $K$ IoT terminals, which completes the proof. □

## 5 Simulation results

In this section, Monte-Carlo simulation is conducted to evaluate the performance of our designed joint offloading decision and resource allocation algorithm for mission-critical IoT enabled MEC systems. We assume that the IoT devices are randomly and uniformly distributed within a disk with radius setting to be 250 m. In total, we have three APs, i.e., $N = 3$. In addition, we assume that the data size of each device is randomly selected from a given interval, namely, $\mathbb{U}_k \in [60, 100]$ bytes for $k \in \mathcal{K}$. Meanwhile, we stipulate that the output size for the task of IoT device $k$ to be $\mathbb{D}_k \in [30, 50]$

**Table 1 Simulation parameters**

| Parameters | Value |
|---|---|
| Cell radius | 250 m |
| IoT device number, $K$ | [4, 24] |
| Number of APs, $N$ | 3 |
| Data volume, $\mathbb{U}_k$ | $\mathbb{U}_k \in [60, 100]$ bytes |
| Output data size, $\mathbb{D}_k$ | $\mathbb{D}_k \in [30, 50]$ bytes |
| Required computation resource $\mathbb{F}_k$ | $\mathbb{F}_k \in [1 \times 10^3, 1 \times 10^4]$ CPU cycles |
| AP computation capacity, $L_n$ | 1 G CPU cycles per second |
| Effective switched capacitance, $\boldsymbol{\alpha}_n$ | $10^{-27}$ |
| Short packet transmission time, $\tau_{\mathrm{U}}$, $\tau_{\mathrm{D}}$ | 20 ms |
| Latency requirement of device $k$, $T_k$, | 5 s |
| SNR loss, $\delta$ | 1.5 |
| Decoding error probability, $\varepsilon_k$ | $2 \times 10^{-8}$ |
| IoT devices distribution | Randomly uniform distribution |
| APs distribution | Uniformly distributed over the circle diameter |
| Noise power spectral density, $N_0$ | − 130 dBm/Hz |
| System bandwidth, $W$ | 1 MHz |
| AP transmit power budget, $P_n$ | 1 W |
| Small scale fading | Rayleigh fading with unit variance |
| Distance dependent path loss | $128.1 + 37.6 \log_{10} d$, $d$ is in Km |

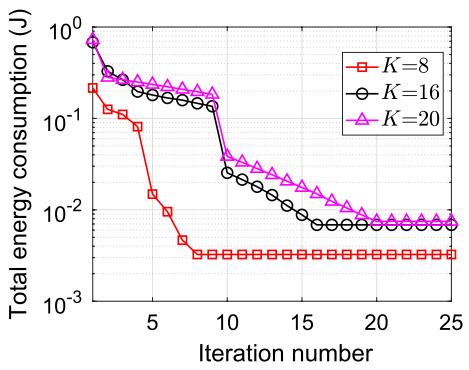


**Fig. 2** Convergence performance

bytes, where $k \in \mathcal{K}$. Besides, the computation capacity per AP is set to be 1 G CPU cycles per second. Moreover, the latency requirement of device $k$, referred to as $T_k$, is assumed to be 5 s. The SNR loss and the noise power spectral density are set as $\delta = 1.5$ and $N_0 = -130$ dBm/Hz, respectively. Furthermore, we assume the system bandwidth and the transmit power per AP to be $W = 1$ MHz and $P_n = 1$ watt, respectively. At last, we declare that the applied small scale fading follows the Rayleigh fading with unit variance. Meanwhile, the distance dependent path loss is expressed as $128.1 + 37.6 \log_{10} d$, in which $d$ is in Km and represents the Euclidean distance between IoT devices and the APs. For brevity, we summarize the simulation parameters in Table 1.

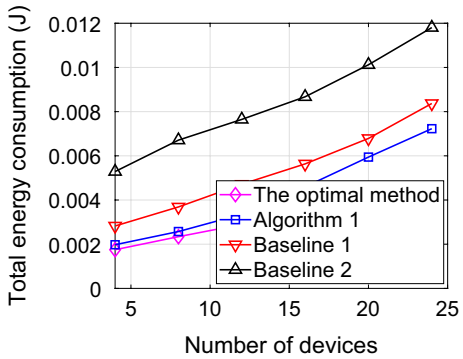


**Fig. 3** The total energy consumption v.s. IoT devices number, $K$

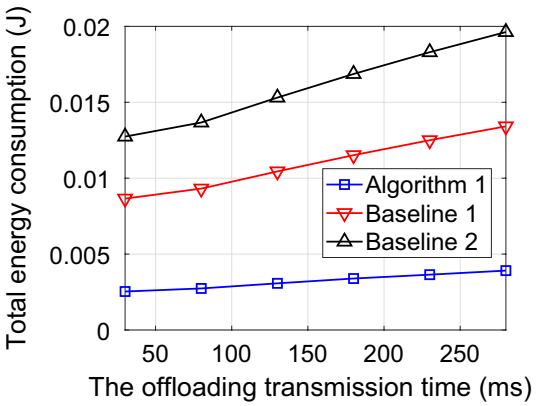


**Fig. 4** The total energy consumption v.s. the offloading transmission time, $\tau_U$

Two baseline schemes are taken into account for performance comparison, and they are listed below:

- Baseline 1: In this scheme, different IoT devices select the nearest AP to do the task offloading.
- Baseline 2: In this strategy, each device chooses its associated AP randomly.

Note that in foregoing mentioned two benchmark schemes, the resource management strategy is identical to that used in our explored scheme. In addition, to provide a comprehensive evaluation, a pair of system performance metrics are considered in this section, i.e., the convergence performance and the total energy consumption, respectively.

Figure 2 shows the convergence performance of our proposed joint offloading decision and resource allocation algorithm, i.e., Algorithm 1. We use the system energy consumption during different iterations to represent this performance. It can be seen from Fig. 2 that, for any given $K$, the designed method can converge rapidly, illustrating the efficiency of our joint optimization approach. In addition, as expected, a larger number of IoT devices, namely, $K$, induces a higher total energy consumption.

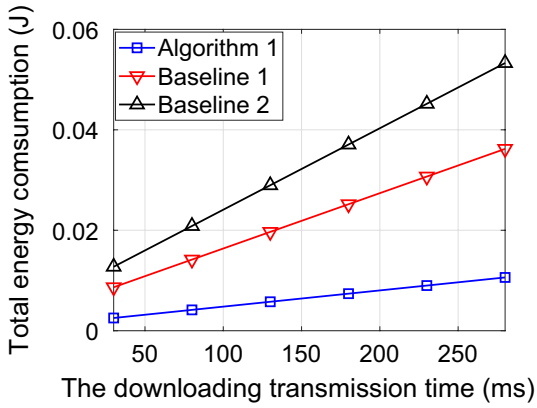


**Fig. 5** The total energy consumption versus the downloading transmission time, $\tau_{\mathrm{D}}$

Figure 3 depicts the energy consumption of four schemes with varying number of devices, where the $x$-axis is the number of IoT devices, while the $y$-axis expresses the total energy consumption. We claim that the optimal method in Fig. 3 is the exhaustive search oriented approach as stated in Sect. 4.1. Due to the limited workspace capacity of MATLAB, the values of the optimal scheme when $K$ is larger than 12 are omitted. Each point on these two figures is obtained via averaging over 300 feasible instances. In addition, two baselines as mentioned before are taken into account to do the comparison with our devised algorithm. From Fig. 3, we can observe that, the energy consumption gap between Algorithm 1 and the optimal scheme is tiny, showing the energy efficiency of our proposed joint optimization approach. In addition, for any given $K$, Algorithm 1 outperforms the other two baselines, demonstrating the significance of offloading decision optimization in terms of saving total energy costs. More specific, when $K = 24$, Algorithm 1 saves energy by 15.8% and 63.3% when compared to Baseline 1 and Baseline 2, respectively. Moreover, it is worth noting that for any given $K$, Baseline 1 requires a less total energy than that of Baseline 2.

Figure 4 expresses the energy consumption of our devised algorithm compared to two baseline schemes, wherein the $x$-axis depicts the offloading time of the IoT devices, referred to as $\tau_{\mathrm{U}}$, while the $y$-axis represents the total energy cost of each approach. In addition, it is stipulated that $\tau_{\mathrm{D}} = 30$ ms. From which, we see that the explored algorithm outperforms the benchmark strategies significantly, especially when $\tau_{\mathrm{U}}$ is large. For instance, when $\tau_{\mathrm{U}} = 280$ ms, our developed method saves energy by 70.9% and 80.1% compared to Baseline 1 and Baseline 2, respectively. Unsurprisingly, for any given $\tau_{\mathrm{U}}$, Baseline 1 always needs less total energy compared to Baseline 2.

At last, we investigate the effect of the downloading transmission time, i.e., $\tau_{\mathrm{D}}$, on the total energy consumption of each schemes, as illustrated in Fig. 5. Therein, the total offloading time is set to be $\tau_{\mathrm{U}} = 30$ ms. As expected, the proposed joint optimization strategy has the best energy efficiency among all the three schemes due to the sophisticatedly designed offloading decision as well as the resource management. Details are not repeated here to avoid redundancy. Comparing Fig. 4 with Fig. 5, we observe that increase the download time of all the IoT devices results in a fast-growing total energy consumption, i.e., the slope of the curves in Fig. 5 is larger than that of Fig. 4. This is

because the data sizes of computing results at APs are smaller than that of the offloaded tasks. Increasing the transmission time for downloading phase degrades the time for tasks' offloading and computing as the total latency, i.e., $T_k$, is fixed for IoT device $k$, where $k \in \mathcal{K}$.

## 6 Conclusion and future work

In this paper, the energy cost minimization for mission-critical IoT in MEC system was studied. The formulated minimization problem is a MINLP problem, which has been decomposed into two sub-problems for analytical tractability, i.e., the task offloading decision-making sub-problem and the resource optimization sub-problem. We demonstrated that the resource management sub-problem of IoT devices under the fixed offloading strategy are convex and can be best solved by the existing convex optimization tools. On this basis, we showed that the optimal solution to the original optimization problem was attainable by an exhaustive search enabled approach, which suffered an exponentially increased time complexity. Thereafter, a time-efficient sub-optimal algorithm was designed, which was conducted in an iterative manner and had a provable convergence guarantee compared with a wide range of benchmark schemes, the simulation results show the effectiveness of our algorithm in terms of convergence and energy consumption. In the future, we will investigate the joint offloading decision and radio resource management for MEC-enabled IoT networks with ultra-low latency consideration. Besides, the optimization for AP deployment will be another important direction.

**Abbreviations**

MEC: Mobile edge computing; IoT: Internet-of-things; AP: Access points; MINLP: Mixed-integer non-linear programming; UAV: Unmanned aerial vehicle; OFDMA: Orthogonal frequency division multiple access; NOMA: Non-orthogonal multiple access; URLLC: Ultra-reliable low-latency communication; eMBB: Enhanced mobile broadband; CSI: Channel state information; SNR: Signal-to-noise-ratio; w.r.t.: With respect to.

**Authors' contributions**

The main idea and the technical analysis of this paper were proposed and designed by Dr. Yaru Fu and Prof. Tony Q. S. Quek. The simulation was done by Dr. Xiaolong Yang. The other authors helped to improve the content as well as presentation of this paper. All authors read and approved the final manuscript.

**Funding**

Not applicable.

**Availability of data and materials**

The data set of this paper is not open for public.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

[1] School of Science and Technology, The Open University of Hong Kong, Hong Kong 999077, China. [2] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372, Singapore. [3] School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China. [4] School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China. [5] School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. [6] National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China.

Received: 5 November 2020 Accepted: 22 January 2021
Published online: 10 February 2021

**References**

1. C. Li, F. Sun, J.M. Cioffi, L. Yang, Energy efficient mimo relay transmissions via joint power allocations. IEEE Trans. Circuits Syst. **61**(7), 531–535 (2014)

2. C. Li, J. Wang, F.-C. Zheng, J.M. Cioffi, L. Yang, Overhearing-based co-operation for two-cell network with asymmetric uplink–downlink traffics. IEEE Trans. Signal Inf. Process. Netw. **2**(3), 350–361 (2016)
3. Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing: a key technology towards 5G. European Telecommunications Standards Institute (ETSI) White Paper (2015)
4. Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, A survey on mobile edge computing: the communication perspective. IEEE Commun. Surv. Tutor. **19**(4), 2322–2358 (2017). (**Fourthquarter**)
5. N. Abbas, Y. Zhang, A. Taherkordi, T. Skeie, Mobile edge computing: a survey. IEEE Internet Things J. **5**(1), 450–465 (2018)
6. T.X. Tran, A. Hajisami, P. Pandey, D. Pompili, Collaborative mobile edge computing in 5g networks: new paradigms, scenarios, and challenges. IEEE Commun. Mag. **55**(4), 54–61 (2017)
7. Y. Du, K. Wang, K. Yang, G. Zhang, Energy-efficient resource allocation in UAV based MEC system for IoT devices. In: *IEEE Global Telecommunications Conference* (Globecom) (2018), pp. 1–6
8. T.Q. Dinh, Q.D. La, T.Q. Quek, H. Shin, Learning for computation offloading in mobile edge computing. IEEE Trans. Commun. **66**(12), 6353–6367 (2018)
9. W. Wen, Y. Fu, T.Q.S. Quek, F.-C. Zheng, S. Jin, Joint uplink/downlink sub-channel, bit and time allocation for multi-access edge computing. IEEE Commun. Lett. **23**(10), 1811–1815 (2019)
10. Z. Yang, Y. Liu, Y. Chen, N. Al-Dhahir, Cache-aided NOMA mobile edge computing: a reinforcement learning approach. IEEE Trans. Wirel. Commun. (2020). https://doi.org/10.1109/TWC.2020.3006922
11. L. Huang, S. Bi, Y.-J.A. Zhang, Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks (2020). arXiv:1808.01977
12. Y.-C. Wu, T.Q. Dinh, Y. Fu, C. Lin, T.Q.S. Quek, A learning-based expected best offloading strategy in wireless edge networks. In: *IEEE Global Telecommunications Conference (Globecom)* (2019), pp. 1–6
13. C. Sun, C. She, C. Yang, T.Q.S. Quek, Y. Li, B. Vucetic, Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications. IEEE Trans. Wirel. Commun. **18**(1), 402–415 (2019)
14. M. Salmani, T.N. Davidson, On multi-user binary computation offloading in the finite-block-length regime. In: *Proceedings 53rd Asilomar Conference on Signals, Systems, and Computers (ACSSC)* (2019), pp. 378–382
15. R. Dong, C. She, W. Hardjawana, Y. Li, B. Vucetic, Deep learning for hybrid 5G services in mobile edge computing systems: learn from a digital twin. IEEE Trans. Wirel. Commun. **18**(10), 4692–4707 (2019)
16. W.R. Ghanem, V. Jamali, R. Schober, Optimal resource allocation for multi-user OFDMA-URLLC MEC systems (2020). arXiv:2009.11073
17. P. Yang, X. Xi, Y. Fu, T.Q.S. Quek, X. Cao, D. Wu, Multicast eMBB and bursty URLLC service multiplexing in a CoMP-enabled RAN (2020). arXiv:2002.09194
18. J. Tang, B. Shim, T.-H. Chang, T.Q. Quek, Incorporating URLLC and multicast eMBB in sliced cloud radio access network. In: *IEEE International Conference on Communications (ICC)* (2019), pp. 1–7
19. A. Anand, G. de Veciana, S. Shakkottai, Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. IEEE/ACM Trans. Netw. **28**(2), 477–490 (2020)
20. Y. Fu, K.N. Doan, T.Q.S. Quek, On recommendation-aware content caching for 6G: an artificial intelligence and optimization empowered paradigm. Digit. Commun. Netw. (2020). https://doi.org/10.1016/j.dcan.2020.06.005
21. M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6G networks: use cases and technologies. IEEE Commun. Mag. **58**(3), 55–61 (2020)
22. Y. Fu, Y. Chen, C.W. Sung, Distributed power control for the downlink of multi-cell NOMA systems. IEEE Trans. Wirel. Commun. **16**(9), 6207–6220 (2017)
23. C. She, C. Yang, Energy efficiency and delay in wireless systems: is their relation always a tradeoff? IEEE Trans. Wirel. Commun. **15**(11), 7215–7228 (2016)
24. W. Yang, G. Durisi, T. Koch, Y. Polyanskiy, Quasi-static multiple antenna fading channels at finite blocklength. IEEE Trans. Inf. Theory **60**(7), 4232–4264 (2014)
25. C. She, C. Liu, T.Q.S. Quek, C. Yang, Y. Li, Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems. IEEE Trans. Commun. **67**(5), 3768–3781 (2019)
26. X. Liu, S. Han, C. Yang, Energy-efficient training-assisted transmission strategies for closed-loop MISO systems. IEEE Trans. Veh. Technol. **64**(7), 2846–2860 (2015)
27. S. Schiessl, J. Gross, H. Al-Zubaidy, Delay analysis for wireless fading channels with finite blocklength channel coding. In: *18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)* (2015), pp. 13–22


## Publisher's Note