Dynamic Resource Allocation and Computation Offloading for IoT Fog Computing System

Zheng Chang, *Senior Member, IEEE,* Liqing Liu, Xijuan Guo, and Quan Sheng

***Abstract*—Fog computing system is able to facilitate computation-intensive applications and emerges as one of the promising technology for realizing the Internet of Things (IoT). By offloading the computational tasks to the fog node (FN) at the network edge, both the service latency and energy consumption can be improved, which is significant for industrial IoT applications. However, the dynamics of computational resource usages in the FN, the radio environment and the energy in the battery of IoT devices make the offloading mechanism design become challenging. Therefore, in this paper, we propose a dynamic optimization scheme for the IoT fog computing system with multiple mobile devices (MDs), where the radio and computational resources, and offloading decisions, can be dynamically coordinated and allocated with the variation of radio resources and computation demands. Specifically, with the objective to minimize the system cost related to latency, energy consumption and weights of MDs, we propose a joint computation offloading and radio resource allocation algorithm based on Lyapunov optimization. Through minimizing the derived upper bound of the Lyapunov drift-plus-penalty function, we divide the main problem into several sub-problems at each time slot and address them accordingly. Through performance evaluation, the effectiveness of the proposed scheme can be verified.**

***Index Terms*—Fog computing; Edge computing; Dynamic computation offloading; Lyapunov optimization; Energy harvesting; Resource allocation**

I. Introduction

A. Background

Nowadays, wireless network is able to provide convenient and reliable connections almost anywhere and anytime. Meanwhile, the emerging Internet of Things (IoT) paradigm boosted by the advanced mobile technologies, is able to provide ubiquitous coverage and information exchange with little human intervention. It is expected that IoT paradigm is able to enable various "smart" applications such as smart city and smart grid. However, due to lack of computational resources and the limitation of battery of the mobile devices (MDs) in the IoT, the gap between demand for complex application and computing capability is gradually increasing. In fact, many latency-sensitive and computational-intensive mobile applications, such as image processing and mobile gaming, may have degraded performance when they are purely executed on MDs [1].

Bearing in mind the aforementioned problems, mobile cloud computing (MCC), including central cloud and fog/edge cloud computing, attracts many attentions recently as one of the solution. In MCC, by offloading computational tasks to the distant cloud and executing them in the cloud, the system energy consumption and latency are able to be improved [2], [3]. In MCC, although the central cloud has huge storage space and rich computational resources, the cloud centers are usually remotely located from the end-users. When connecting with cloud center, the transmission link may be unreliable and also causes long latency. Therefore, the distant cloud center is not desirable for latency-sensitive applications in many cases [4]. Among all different types of MCC technologies, fog/edge computing system, emerges to provide distributed and pervasive computation services for the MDs, and especially for the industrial IoT applications with stringent requirements of latency and reliability [5]- [10]. Fog computing is able to bring both computational and radio resources closer to the MDs, which can improve the scalability from both two perspectives.

Offloading the computation tasks is a promising effective solution for resource-limited MDs to execute the computational-intensive tasks. By offloading the tasks to and receiving results from the computing center, the MDs are able to fully enjoy the complex mobile applications with improved Quality of Experience (QoE), such as service latency experience, and reduced energy consumption. However, for battery-powered MDs, the Quality of Service (QoS), such as throughput or energy consumption of the mobile applications may be degraded due to insufficient energy supply. In addition, due to the limitations of the MDs on size and location, etc, frequent recharging for providing energy supply is not practical in many cases. In this aspect, a promising technology, namely Energy harvesting (EH), which can enable the devices to harvest energy from environment [11], [12] is considered to resolve these issues. By EH from various energy sources, the life time of the MDs can be prolonged and self-sustaining can be expected [13], [14].

As the computing capability of the fog node (FN) is not comparable to the traditional cloud center and one FN only serves a relative small area where the radio resource is also limited, the offloading decisions of the MDs may have a significant impact on the QoS. Accordingly, the usage of the radio resources, such as transmit power and frequency spectrum, and the harvested energy should be carefully designed and optimized in line with the offloading decisions. In addition, as the radio environment and the demand for computational resources

Z. Chang is with School of Computer Science, University of Electronic Science and Technology of China, Chengdu, China, and also with Faculty of Information Technology, University of Jyväskylä, P.O.Box 35, FIN-40014 Jyväskylä, Finland. L. Liu is with Northeasten University, Qinhuangdao 066004, China. X. Guo are with College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China. Q. Sheng is with . This work is supported by NSF of Hebei (No. E2017203351) and Key Research and Development Project of Hebei (No. 19252106D).

vary in a fast speed, dynamic scheduling and optimization are more preferred compared to static optimization schemes. However, because of randomness of radio environment, energy harvesting process and computation demands, realizing the dynamic optimization is challenging. In this paper, the main objective is to overcome the obstacles and provide dynamic resource allocation and computation offloading schemes for EH-enabled IoT fog computing system.

B. Related Work

Most of the researches on the offloading problem concentrate on designing different and effective static schemes for traditional MDs through optimizing the MD's execution decision, radio resource, and/or computational resource [2]– [10]. Considering a fog computing system, the authors of [5] apply queuing theory to investigate the delay, energy consumption, and payment cost of offloading process. Based on the derived theoretical expressions, the authors formulate a multi-objective optimization problem minimizing the cost functions. The problems are addressed by finding the offloading decision and power allocation. In [7], the authors explore the tradeoff between delay and energy consumption in the fog-cloud hybrid computing system. The associated workload allocation problem is addressed accordingly. In [9], the authors propose an optimization framework of offloading to optimize the task allocation decision and the computational resource allocation.

Meanwhile, EH is considered to provide an energy efficient solution for the communication systems due to its self-sustainable nature [11]– [13]. Investigating the impact of EH in cloud computing system has also attracted many interests. In [11], a local data dissemination architecture is investigated combining social networking with EH characteristics. The authors of [12] consider a wireless power transfer (WPT)-aided MCC system, and present a solution for computational and resource allocation to enable computation offloading in such a system. In [13], the authors utilize the Markov chain analysis, and present the design of EH D2D network by modeling the status of harvested energy. It is also worth mentioning that the research of dynamic computation offloading in a fog computing with EH devices does not receive equal attention. The authors of [14]– [19] have studied the WPT-aid edge computing system. In [14], the authors study an edge computing system with EH technologies and investigate the computation offloading problem. The dynamic computation offloading algorithm is proposed to address the formulated problem. In [15], the authors propose nonorthogonal multiple access (NOMA)-enabled computation offloading scheme to minimize the delay, where the MTs offload computation requests to an edge node based on the NOMA transmission. The authors of [16] and [17] the authors investigate the problem of energy consumption and energy-delay tradeoff in an MEC system with multiple EH devices. In [18], the authors utilize the learning-based scheme for addressing the dynamic offloading problems in edge computing system with EH devices. In [18], the authors utilize the learning-based scheme for addressing the dynamic offloading problems in edge computing system with EH devices. In [19], the authors consider unmanned aerial vehicle (UAV)-enabled MEC WPT system and propose efficient scheme to maximize the computation rate.

C. Contribution

As the decisions will be coupled in different time slots, it is challenging to design the resource allocation and computation offloading policies in the considered fog computing system. Inspired by the aforementioned observations, we aim to introduce a dynamic subcarrier allocation, power allocation and computation offloading scheme to minimize the system execution cost via Lyapunov optimization. In order to address confronted challenges, major contributions of this work are summarized as follows:

1) We consider a general fog computing system with multiple MDs equipped with EH capability, an Access Point (AP) and a FN. In particular, we consider different queue models to provide thorough analysis on the delay and energy consumption performance. At the FN, a $M/G/1$ queue is assumed and at the MD, a $M/M/1$ queue is considered.
2) With the derived analytical results, we are able to formulate the system cost consisting of service latency, energy consumption, and the weight/priority of each MD. With the objective to minimize the formulated system cost, the offloading strategy, the transmit power, and the subcarrier assignment are jointly optimized in the proposed resource allocation and offloading scheme.
3) Due to the stochastic nature of the request arrival, the amount of harvested energy and the radio channel, we propose to leverage the advantages of Lyapunov optimization to design an online dynamic algorithm. By minimizing the upper bound of the Lyapunov drift-plus-penalty function from the perspective of different decision variables, the initial problem is divided into several simple sub-problems with low-complexity and addressed accordingly.

D. Organization

The reminder of this paper is organized as follows. We present the system model in Sec. II. In Sec. III the problem is formulated. In Sec. IV, we propose to leverage the advantages of Lyapunov optimization to propose dynamic algorithm to address the formulated problem. The simulation study is conducted in Sec. V. Sec. VI concludes the paper.

II. System Model

The fog computing system consists of N single-core MDs, one AP and one FN, as presented in Fig. 1. The set of MDs is denoted as $\mathcal{N} = \{1, 2, \cdots, N\}$. For executing an application, each MD has a series of homogeneous service requests. At each MD, a first-in-first-out (FIFO) queue is considered for storing arriving requests, and the radio interface is used to establish the wireless connection. As a single processor is assumed, the process queue at the MD is assumed as a $M/M/1$ queue [25]. The MDs are with EH capability which

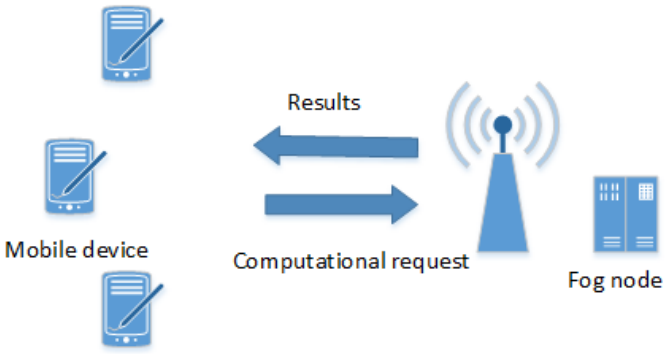


Fig. 1. IoT fog computing system model

TABLE I
SUMMARY OF THE KEY NOTATIONS

Notations	Meanings
$\mathcal{N}$	the set of MDs in the system
N	the total number of MDs in the system
$\mathcal{T}$	the set of time slots
T	the total number of time slots in the system
τ	the length of each time slot
$\mathcal{K}$	the set of subcarriers
K	the total number of subcarriers
$A_i(t)$	the requests arrival rate for MD i at time slot t
θ_i	the data size of request for MD i
$l_i^M(t)$	the workload of MD i at time slot t
u_i^M	the computing capability of MD i
$f_i(t)$	the CPU-cycle frequency of MD i at time slot t
k_i	the switched capacitance of MD i
W	the subcarrier bandwidth
N_0	the noise power spectral density at the receiver of the FN
$h_{i,k}(t)$	the channel gain for MD i on subcarrier k at time slot t
$p_{i,k}(t)$	the transmission power of MD i on subcarrier k at time slot t
$p_{i,max}$	the maximum transmission power of MD i
$\rho_{i,k}(t)$	the subcarrier assignment indicator for MD i at time slot t
u^F	the service rate in the FN
$l^F(t)$	the workload of the FN at time slot t
$e_i(t)$	the harvested energy for MD i at time slot t
$e_i^{\max}(t)$	the maximum harvested energy for MD i at time slot t
$B_i(t)$	the total energy for MD i at time slot t
μ_i	the punishment for per dropped task for MD i
ω_i	the weight factor of MD i
α_i	the weight of task dropping punishment for MD i

enables the MD to obtain energy supply from the environment, and the harvested energy is used for local task execution and wireless data transmission. The AP is responsible for receiving requests from the MD and delivering the requests to the FN for further processing. The process queue of FN is modelled as a $M/G/1$ queue [25]. The MD offloads (part of) the computation requests to the FN to enjoy a higher level of quality of computation experience. The time is slotted and the length of each time slot is τ. The time slot set is denoted as $\mathcal{T} = \{0, 1 \cdots, t, \cdots, T-1\}$. To improve the readability, Table 1 list the key notations.

A. Local Execution Model

The computation requests generated by MD i, $i \in \mathcal{N}$ are assumed to follow Poisson process with an average arrival rate $A_i(t)$ and within $[A_{i,\min}, A_{i,\max}]$. Each request is of data size θ_i. For MD i, some of the computation requests may be locally executed and the rest will be offloaded to the FN. It is worth mentioning that some of the computation requests have to be dropped when neither of these modes is feasible. The decision of MD i at time slot t is modeled as $\Psi_i(t) = \left[\varphi_i^M(t), \varphi_i^F(t), \varphi_i^D(t)\right]$, where $\varphi_i^M(t)+\varphi_i^F(t)+\varphi_i^D(t) = 1$. $\varphi_i^M(t)$ represents the portion that the requests are executed at the MD i at time slot t, $\varphi_i^F(t)$ denotes the portion that the requests are offloaded to the FN, and $\varphi_i^D(t)$ expresses the portion that the requests are dropped.

We denote u_i^M as the computing capability of MD i which depends on CPU Cycle of the MD. Moreover, $l_i^M(t)$ denotes the normalized workload on the MD i at time slot t. For example, $l_i^M(t) = 0$ indicates at time slot t, the CPU is totally idle. Then, the average response time $D_i^M(t)$ for local execution of MD i at time slot t is expressed as follows [20]:

$$D_i^M(t) = \frac{1}{u_i^M\left(1 - l_i^M(t)\right) - \varphi_i^M(t) A_i(t)}. \tag{1}$$

Assume that the available computing capability of MD i is $u_i^M\left(1 - l_i^M(t)\right)$ and the corresponding CPU-cycle frequency is denoted as $f_i(t)$ at time slot t. As shown in [12], under the assumption of a low CPU voltage, the power consumption of CPU is $k_i f_i^3$, where k_i is a constant depending on the switched capacitance of MD, and f_i is the CPU-cycle frequency. Thus, the energy consumption $E_i^M(t)$ of MD i for local execution is given as

$$\begin{aligned} E_i^M(t) =& k_i f_i^3(t) D_i^M(t) \\ =& \frac{k_i f_i^3(t)}{u_i^M\left(1 - l_i^M(t)\right) - \varphi_i^M(t) A_i(t)}. \end{aligned} \tag{2}$$

Nevertheless, if some of the requests cannot be executed, they have to be dropped. We define a cost coefficient μ_i for the task drop, and accordingly the punishment for MD i at time slot t is denoted as

$$C_i^D(t) = \mu_i \varphi_i^D(t) A_i(t) \tau. \tag{3}$$

B. Uplink Transmission

The wireless network is assumed to be Orthogonal Frequency Division Multiplexing (OFDM)-based. The set of the subcarrier is denoted as $\mathcal{K}=\{1, 2 \cdots, k, \cdots, K\}$, where $|\mathcal{K}| = K$. We assume the channels are independent and identically distributed block fading during time slots. Let W denotes the channel bandwidth, N_0 denotes the noise power spectral density at the AP, $h_{i,k}(t)$ denotes the channel gain and $p_{i,k}(t)$ denotes the transmit power of MD i on subcarrier k at time slot t which cannot exceed its maximum value of $p_{i,\max}$. We define $\rho_{i,k}(t) \in \{0, 1\}$ as the subcarrier assignment indicator, where $\rho_{i,k}(t) = 1$ indicates that the subcarrier k is used by MD i at time slot t. Otherwise, $\rho_{i,k}(t) = 0$. In this work, we consider one subcarrier can only be allocated to one MD. Therefore, there is no interference effect among the MDs. Correspondingly, the data rate $r_{i,k}(t)$ of MD i on subcarrier k in the uplink is expressed as follows:

$$r_{i,k}(t) = \rho_{i,k}(t) W \log_2\left(1 + \frac{p_{i,k}(t) h_{i,k}(t)}{N_0 W}\right). \tag{4}$$

In this work, to avoid transmission interference, we consider one subcarrier can only be assigned to one MD , while one MD can be assigned several subcarriers. The total uplink data rate of MD i is denoted as follows:

$$R_i(t) = \sum_{k\in\mathcal{K}} \rho_{i,k}(t)\, W \log_2\left(1 + \frac{p_{i,k}(t)\, h_{i,k}(t)}{N_0 W}\right). \quad (5)$$

Correspondingly, the uplink transmission time $D_i^{up}(t)$ is given as

$$\begin{aligned} D_i^{up}(t) =& \frac{p_i^F(t)\, A_i(t)\, \theta_i \tau}{R_{i,k}(t)} \\ =& \frac{\varphi_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, W \log_2\left(1 + \frac{p_{i,k}(t) h_{i,k}(t)}{N_0 W}\right)}. \end{aligned} \quad (6)$$

Then, the energy consumption $E_i^{up}(t)$ of the transmission is expressed as

$$\begin{aligned} E_i^{up}(t) =& \sum_{k\in\mathcal{K}} \rho_{i,k}(t)\, p_{i,k}(t) D_i^{up}(t) \\ =& \sum_{k\in\mathcal{K}} \frac{\rho_{i,k}(t)\, p_{i,k}(t)\, \varphi_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, W \log_2\left(1 + \frac{p_{i,k}(t) h_{i,k}(t)}{N_0 W}\right)}. \end{aligned} \quad (7)$$

C. Fog Execution Model

The FN connecting to the AP can process the offloaded requests and execute the computation task. We consider the connection between the FN and AP is fiber-based with large enough bandwidth and the transmission time from the AP to FN is ignored. We denote the service rate of the FN as u^F. The pending requests of the MDs are pooled together with a total rate $A_{total}(t)$. Therefore, $A_{total}(t)$ is given as follows:

$$A_{total}(t) = \sum_{i\in\mathcal{N}} \varphi_i^F(t)\, A_i(t). \quad (8)$$

The normalized workload of the FN is denoted as $l^F(t)$, and it presents the occupied portion of each server and $l^F(t) < 1$. The average response time $D^F(t)$ is [21]

$$D^F(t) = \frac{2u^F\left(1 - l^F(t)\right) - \left(\sum\limits_{i\in\mathcal{N}} A_i(t)\ \varphi_i^F(t)\right)}{2u^F\left(1 - l^F(t)\right)\left[u^F\left(1 - l^F(t)\right) - \left(\sum\limits_{i\in\mathcal{N}} A_i(t)\ \varphi_i^F(t)\right)\right]}. \quad (9)$$

After the task is executed at FN, the obtained result is sent to the MDs via AP. Similarly to [6], [7], the energy consumption of the MDs for receiving the results are neglected. This is mainly due to the fact that the data size of the outcome is generally smaller than the one of input.

D. Energy Harvesting Model

A successive energy packet arrival model is used for analyzing the EH process. The arrival of energy packet follows a Poisson process with an average arrival rate $e_i(t)$, and $0 < e_i(t) \leq e_i^{\max}(t)$ where $e_i^{\max}(t)$ is the maximum energy arrival rate at time slot t. The harvested energy is stored in the battery and will be available for further actions. We denote the energy level of the battery of MD i at the beginning of time slot t as $\hat{B}_i(t)$. For simplicity, we only consider the energy consumption for local computation and transmission. The energy consumption $E_{i,total}(t)$ of MD i consists of two parts:

$$E_{i,total}(t) = E_i^M(t) + E_i^{up}(t), \quad (10)$$

where $E_i^M(t)$ is the energy consumption for local processing and $E_i^{up}(t)$ is energy consumption for delivering the requests. Note that $E_{i,total}(t)$ cannot exceed the battery level, i.e.,

$$E_{i,total}(t) \leq \hat{B}_i(t). \quad (11)$$

The energy harvested at time slot t should be used for the next time slot $t+1$, i.e., the battery evolves as follows,

$$\hat{B}_i(t+1) = \hat{B}_i(t) - E_{i,total}(t) + e_i(t). \quad (12)$$

III. Problem Formulation

With the above analysis, the total execution cost compromises of the execution time and the punishment cost for task dropping. The execution time $D_i(t)$ at time slot t is derived as follows:

$$D_i(t) = \varphi_i^M(t)\, D_i^M(t) + \varphi_i^F(t)\left(D_i^{up}(t) + D^F(t)\right). \quad (13)$$

Consequently, the execution cost for MD i can be formulated as

$$EC_i(t) = D_i(t) + \alpha_i C_i^D(t), \quad (14)$$

where α_i is the weight factor of dropping cost. The total weighted execution cost of the system is denoted as $\Gamma_{total}(t)$, which is given as

$$\begin{aligned} &\Gamma_{total}(t) \\ &= \sum_{i\in\mathcal{N}} \omega_i \left[\varphi_i^M(t)\, D_i^M(t) + \varphi_i^F(t)\left(D_i^{up}(t) + D^F(t)\right) + \alpha_i C_i^D(t)\right], \end{aligned} \quad (15)$$

where ω_i is the weight factor, which reflects the relative importance of MD i in the set. Then we derive the average execution cost $\Phi(t)$ of the fog computing system during T time slots, which is given in (16).

At time slot t, the system decision is denoted as $\mathbf{V}(t) = \left[\boldsymbol{\varphi}(t), \boldsymbol{\rho}(t), \mathbf{p}_{up}(t)\right]$, $\forall t \in \mathcal{T}$, where $\boldsymbol{\varphi}(t) = [\boldsymbol{\varphi}_1(t), \cdots, \boldsymbol{\varphi}_i(t), \cdots, \boldsymbol{\varphi}_N(t)]$ are execution strategies for all the MDs at time slot t and $\boldsymbol{\varphi}_i(t) = \left[\varphi_i^M(t), \varphi_i^F(t), \varphi_i^D(t)\right]$ is the execution strategy for MD i at time slot t. $\boldsymbol{\rho}(t) = [\boldsymbol{\rho}_1(t), \cdots, \boldsymbol{\rho}_i(t), \cdots, \boldsymbol{\rho}_N(t)]$ is the subcarrier assignment indicator of the MDs at time slot t. $\boldsymbol{\rho}_i(t) = [\rho_{i,1}(t), \cdots, \rho_{i,k}(t), \cdots, \rho_{i,K}(t)]$ is the subcarrier assignment vector for MD i at time slot t. $\boldsymbol{p}_{up}(t) = [\boldsymbol{p}_1(t), \cdots, \boldsymbol{p}_i(t), \cdots, \boldsymbol{p}_N(t)]$ is the uplink transmit power matrix for all the MDs at time slot t and $\boldsymbol{p}_i(t) = [p_{i,1}(t), \cdots, p_{i,k}(t), \cdots, p_{i,K}(t)]$ is the set of transmit power for MD i.

Therefore, the problem is formulated in the following,

$$\min_{\boldsymbol{V}(t)}\ \Phi(t), \quad (17)$$

$$\begin{aligned}\Phi(t) &= \lim_{T\to+\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\Gamma_{total}(t)\\ &= \lim_{T\to+\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{N}}\omega_i\left[\varphi_i^M(t)D_i^M(t)+\varphi_i^F(t)\left(D_i^{up}(t)+D^F(t)\right)+\alpha_i C_i^D(t)\right].\end{aligned} \tag{16}$$

s.t.

$$\varphi_i^M(t)+\varphi_i^F(t)+\varphi_i^D(t)=1, 0\le\varphi_i^M(t),\varphi_i^F(t),\varphi_i^D(t)\le 1; \tag{18a}$$

$$\varphi_i^M(t)A_i(t)-u_i^M\left(1-l_i^M(t)\right)<0; \tag{18b}$$

$$\sum_{i\in\mathcal{N}}\varphi_i^F(t)A_i(t)-u^F\left(1-l^F(t)\right)<0; \tag{18c}$$

$$0<p_{i,k}(t)<p_{i,\max}; \tag{18d}$$

$$\sum_{i\in\mathcal{N}}\rho_{i,k}(t)\le 1,\quad \rho_{i,k}\in\{0,1\}; \tag{18e}$$

$$E_{i,total}(t)\le\hat{B}_i(t); \tag{18f}$$

$$i\in\mathcal{N}, t\in\mathcal{T}, k\in\mathcal{K}. \tag{18g}$$

As we can see, the decisions are coupled among different time slots due to the constraints (18f), which makes the problem hard to be tackled. As presented in [24], by introducing a reasonable upper bound $E_i^{\max}(t)$ and a non-negative lower bound $E_i^{\min}(t)$ of the battery, the coupling effect is eliminated. Correspondingly, we are able to modify the problem as

$$\mathbf{P1}: \min_{\boldsymbol{V}(t)}\ \Phi(t)$$
$$\text{(18a)}-\text{(18e)}, \text{(18g)}, \tag{19}$$

$$E_{i,total}(t)\in\left[E_i^{\min}(t), E_i^{\max}(t)\right]. \tag{20}$$

As proved in [14], when $E_i^{\min}$ approaches 0, the optimal solution of **P1** will be the same as the original problem. For **P1**, a stochastic optimization problem is formulated with decision variables of the execution strategy, the uplink transmit power and the subcarrier assignment. In general, by addressing the deterministic per-time slot problem, we can obtain the total optimal decisions in a stochastic manner.

IV. Proposed Solution

In this section, we will propose a dynamic algorithm based on Lyapunov optimization to solve the formulated problem. Lyapunov optimization is considered as an efficient tool to design online control algorithm without requiring any prior knowledge [22]–[24]. However, as the battery level is temporally correlated, the decision sets in different time slots are not independent and identically distributed. Therefore, we advocate the weighted perturbation method [27] to solve this issue. We first define the perturbation parameter χ, and a virtual energy queue $B_i(t)=\hat{B}_i(t)-\chi$. The perturbation parameter χ_i is a bounded constant satisfying $\chi_i\ge E_{max}+\nu\alpha_i/E_{min}$, where $\hat{E}_i^{max}$ is related to higher bound of the energy E_i^{max} and CPU-cycle [27]. Then, in order to present the proposed solution, the Lyapunov function is defined as follows:

$$L(\boldsymbol{B}(t))=\frac{1}{2}\sum_{i\in\mathcal{N}}{B_i}^2(t), \tag{21}$$

where $\boldsymbol{B}(t)=[B_1(t),\cdots,B_i(t),\cdots B_N(t)]$. Thus, the conditional Lyapunov drift can be expressed as

$$\Delta(\boldsymbol{B}(t))=E\left[L(\boldsymbol{B}(t+1))-L(\boldsymbol{B}(t))\,|\boldsymbol{B}(t)\right]. \tag{22}$$

The Lyapunov drift-plus-penalty function can be given as follows:

$$\Delta_\nu(\boldsymbol{B}(t))=\Delta(\boldsymbol{B}(t))+\nu E\left[\Gamma_{total}(t)\,|\boldsymbol{B}(t)\right], \tag{23}$$

where $\nu\in(0,+\infty)$ is a control parameter. Then we will find an upper bound of $\Delta(\boldsymbol{B}(t))$ under any feasible set of $\boldsymbol{V}(t)$, which can be found in the following lemma.

Lemma 1. *For any feasible set of $\boldsymbol{V}(t)$, which satisfies (19) and (20), the Lyapunov drift-plus-penalty function $\Delta_\nu(\boldsymbol{B}(t))$ is upper bounded, i.e.,*

$$\begin{aligned}\Delta_\nu(\boldsymbol{B}(t))\le\kappa&+\sum_{i\in\mathcal{N}}\left\{B_i(t)\left[e_i(t)-E_{i,total}(t)\right]\right\}\\&+\nu E\left[\Gamma_{total}(t)\,|\boldsymbol{B}(t)\right],\end{aligned} \tag{24}$$

where κ is a constant, and it is denoted as

$$\kappa=\sum_{i\in\mathcal{N}}\left[\frac{\left(e_i^{\max}(t)\right)^2+\left(E_i^{\max}(t)\right)^2}{2}\right]. \tag{25}$$

Proof. See the Appendix A. □

The key idea of the proposed algorithm is to minimize the upper bound of $\Delta_\nu(\boldsymbol{B}(t))$ in the right-hand side of (24). The proposed algorithm is displayed in Algorithm 1. In this algorithm, by solving a deterministic per-time slot problem, the formulated problem is able to be addressed.

Due to the high complexity of the the considered problem, in the next subsection, we will divide it into several sub-problems to obtain the optimal system decision.

A. Optimal Offloading Strategy

Firstly, we seek the optimal offloading strategy at each time slot t, while taking the other pending variables as constants. Then the problem is translated into the following sub-problem **SP1**, which is

$$\min_{\boldsymbol{p}(t)}\ \sum_{i\in\mathcal{N}}-B_i(t)E_{i,total}(t)+\nu\sum_{i\in\mathcal{N}}\omega_i\left[D_i(t)+\alpha_i C_i^D(t)\right] \tag{26}$$

Algorithm 1 Proposed online algorithm

Step 1: At the beginning of the time slot t, obtain $\boldsymbol{B}(t)$.
Step 2: Through solving the following problem **P2**, determine the system decision set $\boldsymbol{V}(t) = \left[\boldsymbol{\varphi}(t), \boldsymbol{\rho}(t), \boldsymbol{p}_{up}(t)\right]$ that can minimize **P2**:

$$\min_{\boldsymbol{V}(t)} \sum_{i\in\mathcal{N}} \{B_i(t)\left[e_i(t) - E_{i,total}(t)\right]\} + \nu E\left[\Gamma_{total}(t)\left|\boldsymbol{B}(t)\right|\right]$$

s.t. (19), (20)
Step 3: Sset $t = t+1$, update $\boldsymbol{B}(t)$ according to (21), repeat Step 1 and Step 2, until obtain the system decisions of all the time slots.

s.t.

$$(18a) - (18e), (18g), (20).$$

It can be found that (18c) is a coupled constraint, which includes various offloading decision variables of different MDs. Similarly to the ones in [25], we can formulate the proposed problem as a Generalized Nash Equilibrium Problem (GNEP). The exponential penalty function method is applied to transform the original GNEP into a classical NEP and address it by semi-smooth Newton method with Armijo line search. The semi-smooth Newton algorithm, with strong system computing power, inherits many excellent features from the classic Newton algorithm. Through determining step size of the Newton direction with the linear searc strategy, it avoids the sensitivity of the algorithm to initial values. Thus the local convergence becomes global convergence. The proof can be found in Appendix of [25]. Consequently, in the following, we will address the transmit power and subcarrier allocation problems.

B. Power and Subcarrier Allocation

Similarly, the optimal transmit power $\mathbf{p}_{up}(t)$ and subcarrier assignment matrix $\boldsymbol{\rho}(t)$ can be obtained by solving the following sub-problem **SP2** through removing some irrelevant parameters from **P2**, which is denoted as follows:

$$\min_{\{\boldsymbol{\rho}(t), p_{up}(t)\}} \sum_{i\in\mathcal{N}} -B_i(t) E_i^{up}(t) + \nu \sum_{i\in\mathcal{N}} \omega_i \varphi_i^F(t) D_i^{up}(t), \tag{27}$$

s.t.

$$0 < p_{i,k}(t) < p_{i,\max}; \tag{28a}$$

$$\sum_{i\in\mathcal{N}} \rho_{i,k}(t) \le 1, \quad \rho_{i,k} \in \{0,1\}; \tag{28b}$$

$$E_i^{up}(t) < E_i^{\max}(t). \tag{28c}$$

By substituting the specific expressions of $E_i^{up}(t)$ and $D_i^{up}(t)$ into the above problem, we can get an equal form of **SP2**, as shown in **SP2**$'$.

The constraints are the same as those in (28). It can be found that **SP2**$'$ is a mixed-integer programming problem. To address such a problem, we introduce an average offloading priority function [26], and it is defined as follows:

$$\psi_{i,k,t}(\omega_i, \tau, h_{i,k}(t)) = \begin{cases} \frac{\omega_i N_0 W}{h_{i,k}(t)}\left[v_i(t)\ln v_i(t) - v_i(t) + 1\right], \\ \text{if } v_i(t) \ge 1; \\ 0, \text{if } v_i(t) < 1; \end{cases} \tag{30}$$

where the constant $v_i(t)$ is defined as $v_i(t) = \frac{W h_{i,k}(t)\tau c_0}{N_0 \ln 2}$ and c_0 is a pre-defined constant. Specifically, with the defined average offloading priority function $\psi_{i,k,t}(\omega_i, \tau, h_{i,k}(t))$ (for simplify, we assume that any two values of $\psi_{i,k,t}(\omega_i, \tau, h_{i,k}(t))$ are not the same), we denote the offloading priority order as $\Psi(t)$ at time slot t, which is composed by $\{\psi_{i,k,t}\}, i \in \mathcal{N}, k \in \mathcal{K}$, and displayed in the descending manner. We denote the sets of assigned and unassigned subcarriers as $\mathcal{K}_1(t)$ and $\mathcal{K}_2(t)$ at the beginning of time slot t. The average channel gain $\tilde{h}_i(t)$ is defined as $\tilde{h}_i(t) = \sum_{k\in\mathcal{K}_2(t)} h_{i,k}(t)/|\mathcal{K}_2(t)|$, where $|\mathcal{K}_2(t)|$ is the number of unassigned subcarriers during the time slot t.

For each MD, such as MD i, the assigned subcarrier set is denoted as $\mathcal{Z}_i(t)$ during the time slot t, initialized as $\mathcal{Z}_i(t) = \varnothing$. Additionally, the subcarrier assignment indicators are set as $\{\rho_{i,k}(t) = 0\}$ at the beginning of time slot t. By these definitions, we proposed a subcarrier allocation algorithm, which is shown in Algorithm 2.

In the proposed algorithm, finding the optimal power allocation involves addressing **SP2**$''$ where $n_i(t)$ is the total number of subcarriers that allocated to MD i. We can also find that **SP2**$''$ is a mixed integer programming problem including a integer constraint (32a). The problem can be addressed with semi-smooth Newton method, which is similar to the one in [25]. Then by the branch-and-bound procedure, we can obtain the integer solution $n_i^*(t)$.

We denote the set of MDs that still require subcarriers as $\tilde{\mathcal{N}}$, where $\tilde{\mathcal{N}} = \{i \,|n_i^*(t) > 0\}$. We allocate subcarriers for each MD according to the highest offloading priority principle. After searching for the highest offloading priority $\psi_{i',k',t}$ over

Algorithm 2 Subcarrier assignment algorithm

1: **Input**:
   At beginning of time slot t, obtain $\Psi(t)$, $h_{i,k}(t)$,$\mathcal{K}_1(t)$, $\mathcal{K}_2(t)$, and $\tilde{h}_i(t)$;
2: **Obtain the total number of subcarriers**:
   Obtaining the optimal solution $\{n_i^*(t), \tilde{p}_i^*(t)\}$ of the **SP2**$''$;
3: **Subcarrier allocation**:
4: **while** $\tilde{\mathcal{N}} \neq \varnothing$, **do**
5: (1) Let $\rho_{k',n'} = 1$, where $\{i', k'\} = \arg\max_{i'\in N, k'\in K} \psi_{i,k,t}$;
   (2) Update sets:
   $\mathcal{Z}_{i'}(t) = \mathcal{Z}_{i'}(t) \cup \{k'\}$,$\mathcal{K}_1(t) = \mathcal{K}_1(t) \cup \{k'\}$, $\mathcal{K}_2(t) = \mathcal{K}_2(t) \setminus \{k'\}$;
   (3) if $|\mathcal{Z}_{i'}(t)| = \tilde{n}_{i'}^*(t)$, then $\tilde{\mathcal{N}} = \tilde{\mathcal{N}} \setminus \{i'\}$;
6: **end while**
7: **Transmit power allocation**
   Obtaining the optimal solution of **SP2**$'''$.
8: **return** $\left\{\rho_{i,k}^*(t), p_{i,k}^*(t)\right\}$

$$\mathbf{SP2}': \min_{\{\boldsymbol{\rho}(t),\boldsymbol{p_{up}}(t)\}} \sum_{i\in\mathcal{N}} -B_i(t) \sum_{k\in\mathcal{K}} \frac{\rho_{i,k}(t)\, p_{i,k}(t)\, \varphi_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, B\log_2\left(1+\frac{p_{i,k}(t) h_{i,k}(t)}{N_0 W}\right)} + \nu \sum_{i\in\mathcal{N}} \omega_i \varphi_i^F(t) \left(\frac{\varphi_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, W\log_2\left(1+\frac{p_{i,k}(t) h_{i,k}(t)}{N_0 W}\right)} \right) \tag{29}$$

$$\mathbf{SP2}'': \min_{\{\boldsymbol{n}_i(t),\tilde{\boldsymbol{p}}_i(t)\}} \sum_{i\in\mathcal{N}} \left\{ \frac{-B_i(t)\,\tilde{p}_i(t)\,\varphi_i^F(t)\,A_i(t)\,\theta_i\tau}{B\log_2\left(1+\frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 W}\right)} + \frac{\nu\omega_i\left[\varphi_i^F(t)\right]^2 A_i(t)\,\theta_i\tau}{n_i(t)\,W\log_2\left(1+\frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 W}\right)} \right\} \tag{31}$$

s.t.

$$\sum_{i\in\mathcal{N}} n_i(t) \le |\mathcal{K}_2(t)| \tag{32a}$$

$$\tilde{p}_i(t) \le p_{i,\max} \tag{32b}$$

$$\frac{\tilde{p}_i(t)\,\varphi_i^F(t)\,A_i(t)\,\theta_i\tau}{W\log_2\left(1+\frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 W}\right)} \le E_i^{\max}(t) \tag{32c}$$

unassigned subcarriers $\mathcal{K}_2(t)$ for the remaining offloading-required users $\tilde{\mathcal{N}}$ and then allocates subcarrier k' to user i'. Such a sequential subcarrier assignment follows the descending offloading priority order. Then the remaining sets can be updated until all subcarriers are assigned. At last, the optimal transmit power for MD i over the assigned subcarriers $\mathcal{Z}_i(t)$ is obtained by minimizing the problem $\mathbf{SP2}'''$ at time slot t, i.e.

$\mathbf{SP2}'''$:

$$\min_{p_{i,k'}(t),k'\in\mathcal{Z}_i(t)} \sum_{k'\in\mathcal{Z}_i(t)} \frac{-B_i(t)\,p_{i,k'}(t)\,\varphi_i^F(t)\,A_i(t)\,\theta_i\tau}{\sum\limits_{k'\in\mathcal{Z}_i(t)} W\log_2\left(1+\frac{p_{i,k'}(t)h_{i,k'}(t)}{N_0 W}\right)} + \frac{\nu\omega_i\left(\varphi_i^F(t)\right)^2 A_i(t)\,\theta_i\tau}{\sum\limits_{k'\in\mathcal{Z}_i(t)} W\log_2\left(1+\frac{p_{i,k'}(t)h_{i,k'}(t)}{N_0 W}\right)}, \tag{33}$$

s.t.

$$0 < p_{i,k'}(t) \le p_{i,\max}, k' \in \mathcal{Z}_i(t), \tag{34a}$$

$$\sum_{k'\in\mathcal{Z}_i(t)} \frac{p_{i,k'}(t)\,\varphi_i^F(t)\,A_i(t)\,\theta_i\tau}{\sum\limits_{k'\in\mathcal{Z}_i(t)} W\log_2\left(1+\frac{p_{i,k'}(t)h_{i,k'}(t)}{N_0 W}\right)} \le E_i^{\max}(t), \tag{34b}$$

which equals to

$$\min_{p_{i,k'}(t),k'\in\mathcal{Z}_i(t)} \frac{-c_1\left(\sum\limits_{k'\in\mathcal{Z}_i(t)} p_{i,k'}(t)\right)+c_2}{\sum\limits_{k'\in\mathcal{Z}_i(t)} W\log_2\left(1+\frac{p_{i,k'}(t)h_{i,k'}(t)}{N_0 W}\right)} \tag{35}$$

where c_1 and c_2 are constants, which are denoted as follows:

$$\begin{aligned} c_1 &= B_i(t)\varphi_i^F(t)\,A_i(t)\,\theta_i\tau, \\ c_2 &= \nu\omega_i\left(\varphi_i^F(t)\right)^2 A_i(t)\,\theta_i\tau \end{aligned} \tag{36}$$

We can see that the formulated problem $\mathbf{SP2}'''$ is similar with the problem investigated in [5]. Then, we can solve it with Interior Point Method (IPM), the details of which can be found in [5] (Alg. 1 in Sec. VI).

The presented solution involves two subproblems, namely, the offloading strategy and the radio resource (subcarrier and power) allocation, and they are hierarchically interconnected. Convergence is guaranteed since in each sublayer, the presented solution is able to obtain optimal solution with guaranteed convergence, respectively.

V. Performance Evaluations

In this section, we have conducted simulations to examine and illustrate the proposed resource allocation and computation offloading scheme The simulation parameters are similar to the ones used in [5] and [25].

First, we change the number of subcarriers and plot the average execution cost assuming 6 MDs in Fig. 2. We also compare our propose scheme with the other two schemes, namely equal allocation and random allocation, to show the impact of the proposed subcarrier allocation scheme. It can be observed that with the proposed optimal subcarrier allocation strategy, the average execution cost of the system is the smallest among all three schemes. Moreover, as shown in this figure, when the numebr of subcarriers in the system increases, the average execution cost becomes smaller, as the MDs have sufficient choices to offload the requests to the FN to reduce the execution time. In this way, the number of dropped requests would also be reduced.

Then we plot the execution cost of the system with the increasing number of MDs in the system, when the number of subcarriers is fixed in Fig. 3. First, it can be observed when the number of MDs in the system becomes larger, the

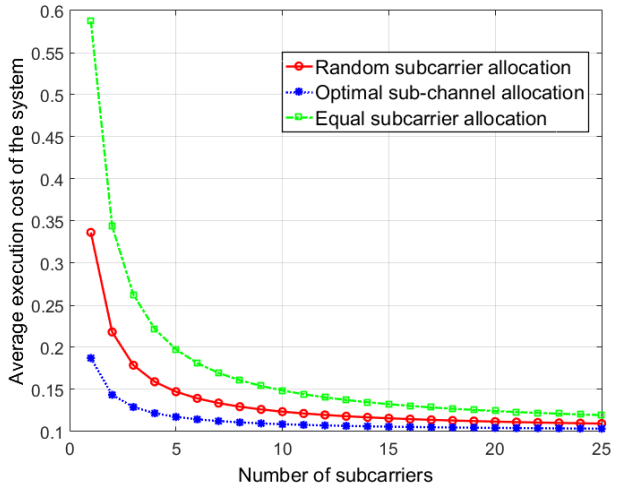


Fig. 2. Execution cost vs. number of subcarriers

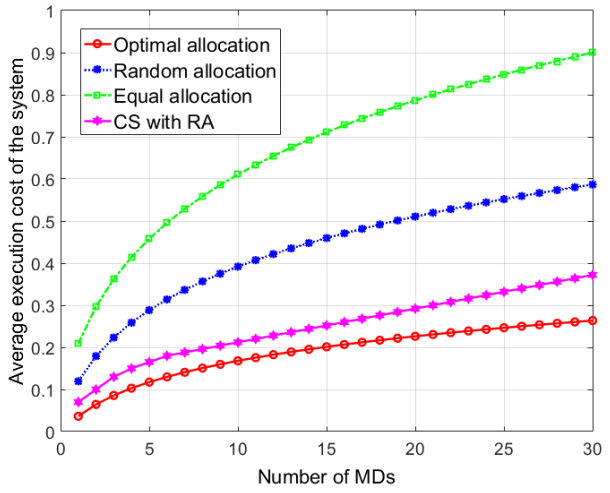


Fig. 3. Execution cost vs. number of MDs

average execution costs are increased. Such a phenomenon indicates that the execution time and/or the punishment cost are degraded with the increasing number of MDs under the condition of fixed number of subcarriers. As more and more users compete for the radio and computational resources with each other, longer transmission time and fog execution time can be induced. Thus, more requests have to be executed locally or even dropped, which leads to a larger execution cost. In addition, we have compared our work with computation resource allocation scheme in [24]. We have implemented the subcarrier allocation but no power allocation. It can be found that our proposed one outperforms all the three schemes.

VI. Conclusion

In this paper, a resource allocation and computation offloading scheme is proposed for fog computing system with multiple EH MDs. Based on Lyapunov optimization, a dynamic algorithm is presented to solve the considered problem over consecutive time slots. To address the formulated problem, we divide the original problem into three sub-problems. Specifically, through transforming the original simplified problem into a GNEP and solving it with semi-smooth Newton algorithm, we can obtain the optimal offloading strategy. Then, we derive the optimal subcarrier assignment scheme, and further obtain the optimal transmit power with IPM algorithm. With such a process, the average execution cost of the fog computing system can be minimized. The performance evaluations are presented to illustrate the effectiveness of the proposed scheme and demonstrate the superior performance over the existing schemes.

As one future research direction, we are going to investigate the interplay of the cloud center and fog node, where the fog node can further offload the data to the cloud center if it does not have enough computational resources. In addition, we will also extend the work to the scenario with multiple fog nodes, where the association between the fog nodes and MDs will be studied.

Appendix A

As we have $B_i(t+1) = B_i(t) - E_{i,total}(t) + e_i(t)$, by squaring both sides, we can obtain:

$$\begin{aligned}
{B_i}^2(t+1) =& [B_i(t) + e_i(t) - E_{i,total}(t)]^2 \\
=& {B_i}^2(t) + [e_i(t) - E_{i,total}(t)]^2 \\
&+ 2B_i(t)[e_i(t) - E_{i,total}(t)] \\
\le& {B_i}^2(t) + {e_i}^2(t) + E_{i,total}^2(t) + 2B_i(t)[e_i(t) - E_{i,total}(t)]
\end{aligned}$$

By moving the expression ${B_i}^2(t)$ to the left-hand side, we can obtain:

$$\begin{aligned}
&{B_i}^2(t+1) - {B_i}^2(t) \\
&\le {e_i}^2(t) + E_{i,total}^2(t) + 2B_i(t)[e_i(t) - E_{total}(t)].
\end{aligned}$$

By summing up the inequalities for $i = 1, 2, \cdots, i, \cdots, N$, we can obtain:

$$\begin{aligned}
&\frac{1}{2}\sum_{i\in\mathcal{N}}\left({B_i}^2(t+1) - {B_i}^2(t)\right) \\
&\le \frac{1}{2}\sum_{i\in\mathcal{N}}\{{e_i}^2(t) + E_{i,total}^2(t) + 2B_i(t)[e_i(t) - E_{total}(t)]\}
\end{aligned}$$

As $0 < e_i(t) \le e_i^{\max}(t)$, $E_{i,total}(t) \le E_{i,\max}(t)$, so we can obtain $e_i^2(t) \le (e_i^{\max}(t))^2$, $E_{i,total}^2(t) \le E_{i,\max}^2(t)$. Thus, we have

$$\begin{aligned}
&\Delta(\boldsymbol{B}(t)) \\
=& E[L(\boldsymbol{B}(t+1)) - L(\boldsymbol{B}(t))\,|\boldsymbol{B}(t)] \\
=& \frac{1}{2}\sum_{i\in\mathcal{N}}\left({B_i}^2(t+1) - {B_i}^2(t)\right) \\
\le& \frac{1}{2}\sum_{i\in\mathcal{N}}\left\{(e_i^{\max}(t))^2 + E_{i,\max}^2(t) + 2B_i(t)[e_i(t) - E_{i,total}(t)]\right\} \\
=& \sum_{i\in\mathcal{N}}\left\{\frac{(e_i^{\max}(t))^2 + E_{i,\max}^2(t)}{2} + B_i(t)[e_i(t) - E_{i,total}(t)]\right\}
\end{aligned}$$

So we can obtain:

$$\begin{aligned}
&\Delta_\nu(\boldsymbol{B}(t)) \\
&= \Delta(\boldsymbol{B}(t)) + \nu E[\Gamma_{total}(t)\,|\boldsymbol{B}(t)] \\
&\le \sum_{i\in\mathcal{N}}\left\{\frac{(e_i^{\max}(t))^2 + E_{i,\max}^2(t)}{2} + B_i(t)[e_i(t) - E_{i,total}(t)]\right\} \\
&\quad + \nu E[\Gamma_{total}(t)\,|\boldsymbol{B}(t)]
\end{aligned}$$

REFERENCES

[1] G. Guerrero-Contreras, J. L. Garrido, S. Balderas-Diaz, and C. Rodriguez-Dominguez, "A context-aware architecture supporting service availability in mobile cloud computing," *IEEE Transactions on Services Computing*, vol. 10, no. 6, pp. 956-968, Nov.-Dec. 2017.

[2] X. Guo, L. Liu, Z. Chang, and T. Ristaniemi, "Data offloading and task allocation for cloudlet-assisted ad hoc mobile clouds," *Wireless Network*, vol. 24, no. 1, pp. 79-88, Jan. 2018.

[3] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Vehicular Technology*, vol. 67, no. 1, pp. 44-55, Jan. 2018.

[4] N,. Zhao, X. Liu, F. R. Yu, M. Li, Victor C. M. Leung, "Communications, caching, and computing oriented small cell networks with interference alignment," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 29-35, Sept. 2016.

[5] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multi-objective optimization for computation offloading in fog computing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 283-294, Feb. 2018.

[6] Y. Sun, S. Zhou and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra-dense networks," *IEEE Journal on Seleted Area in Communications*, vol. 35, no. 11, pp. 2637-2646, Nov. 2017.

[7] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing towards balanced delay and power consumption," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1171-1181, Dec. 2016.

[8] X. Meng, W. Wang, and Z. Zhang, "Delay-constrained hybrid computation offloading with cloud and fog computing," *IEEE Access*, vol. 5, pp. 21355-21367, Sep. 2017.

[9] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, "Offloading in mobile edge computing: task allocation and computational frequency scaling," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3571-3584, Apr. 2017.

[10] Q. Zhu, B. Si, F. Yang, and Y. Ma, "Task offloading decision in fog computing system," *China Communications*, vol. 14, no. 11, pp. 59-68, Dec. 2017.

[11] L. Jiang, H. Tian, Z. Xing, K. Wang, K. Zhang, S. Maharjan, S. Gjessing, and Y. Zhang, "Social-aware energy harvesting device-to-device communications in 5G networks," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 20-27, Aug. 2016.

[12] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE Journal on Selected Areas in Communications*, vol.34, no. 5, pp. 1757-1771, Mar. 2016.

[13] H. H. Yang, J. Lee, and T. Q. S. Quek, "Heterogeneous cellular network with energy harvesting-based D2D communication,"*IEEE Transactions on Wireless Communications*, vol.15, no.2, pp. 1406-1419, Feb. 2016.

[14] Y. Mao, J. Zhang, K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590-3605, Dec. 2016.

[15] Y. Wu, L. P. Qian, K. Ni, C. Zhang and X. Shen, "Delay-Minimization Nonorthogonal Multiple Access Enabled Multi-User Mobile Edge Computation Offloading," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 392-407, June. 2019.

[16] G. Zhang, Y. Chen, Z. Shen and L. Wang, "Distributed Energy Management for Multiuser Mobile-Edge Computing Systems With Energy Harvesting Devices and QoS Constraints," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6606-6617, Aug. 2019.

[17] G. Zhang, W. Zhang, Y. Cao, D. Li and L. Wang, "Energy-Delay Tradeoff for Dynamic Offloading in Mobile-Edge Computing System With Energy Harvesting Devices," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4642-4655, Oct. 2018.

[18] Z. Wei, B. Zhao, J. Su and X. Lu, "Dynamic Edge Computation Offloading for Internet of Things With Energy Harvesting: A Learning Method," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4436-4447, June 2019.

[19] F. Cheng et al., "UAV Trajectory Optimization for Data Offloading at the Edge of Multiple Cells," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6732-6736, July 2018.

[20] A. Lazar, "The throughput time delay function of an M/M/1 queue," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 914-918, Nov. 1983.

[21] R. E. Machol, "Queue theory," *IRE Transactions on Education*, vol. E-5, no. 2, pp. 99-105, Nov. 2007.

[22] Z. Jiang, and S. Mao, "Energy delay tradeoff in cloud offloading for multi-core mobile devices," *IEEE Access*, vol. 3, pp. 2306-2316, Nov. 2015.

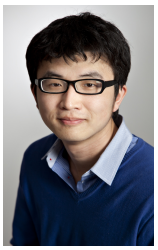
[23] J. Wang, J. Peng, Y. Wei, D. Liu, and J. Fu, "Adaptive application offloading decision and transmission scheduling for mobile cloud computing," *China Communications*, vol. 14, no. 3, pp. 169-181, Mar. 2017.

[24] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590-3605, Dec. 2016.

[25] L. Liu, Z. Chang, and X. Guo, "Socially-aware dynamic computation offloading scheme for fog computing system with energy harvesting devices," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 283-294, Mar. 2018.

[26] C. You, K. Huang, H. Chae, and B. H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397-1411, Mar. 2017.

[27] M. J. Neely and L. Huang, "Dynamic product assembly and inventory control for maximum profit," *in Proc. IEEE Conf. Decision Control (CDC)*, Atlanta, GA, USA, Dec. 2010.



Zheng Chang (S'10-M'13-SM'17) received the B.Eng. degree from Jilin University, Changchun, China in 2007, M.Sc. (Tech.) degree from Helsinki University of Technology (Now Aalto University), Espoo, Finland in 2009 and Ph.D degree from the University of Jyväskylä, Jyväskylä, Finland in 2013. Since 2008, he has held various research positions at Helsinki University of Technology, University of Jyväskylä and Magister Solutions Ltd in Finland. He was a visiting researcher at Tsinghua University, China, from June to August in 2013, and at University of Houston, TX, from April to May in 2015. He has been awarded by the Ulla Tuominen Foundation, the Nokia Foundation and the Riitta and Jorma J. Takanen Foundation for his research excellence. He has been awarded as 2018 IEEE Communications Society best young researcher for Europe, Middle East and Africa Region.

He has published over 100 papers in Journals and Conferences, and received received best paper awards from IEEE TCGCC and APCC in 2017. He serves as an editor of IEEE Access, Springer Wireless Networks and International Journal of Distributed Sensor Networks, and a guest editor for IEEE Network, IEEE Wireless Communications, IEEE Communications Magazine, IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics, Physical Communications, EURASIP Journal on Wireless Communications and Networking, and Wireless Communications and Mobile Computing. He was the exemplary reviewer of IEEE Wireless Communication Letters in 2018. He has participated in organizing workshop and special session in Globecom' 19, WCNC'18-20, SPAWC'19 and ISWCS'18. He also serves as TPC member for many IEEE major conferences, such as INFOCOM, ICC, and Globecom. His research interests include IoT, cloud/edge computing, security and privacy, vehicular networks, and green communications.

Liqing Liu received her master degree in College of Science at Yanshan University in 2015 and PhD degree in College of Information Science and Engineering at Yanshan University, Qinhuangdao, China. She is now with Northeastern University, Qinhuangdao, China. Her research interests include cloud computing and mobile computing.

Xijuan Guo received a PhD degree from Yanshan University. She is now a professor at College of Information Science and Engineering, Yanshan University, Qinhuangdao, China. Her research interests include high performance computing, cloud computing, image processing, wireless communications.

Quan Sheng received his B.S degree in electronics science and technology and Ph.D degree in opto-electronics technology from Tianjin University, Tianjin, China, in 2008 and 2013, respectively. He is currently an associate professor with the School of Precision Instrument and Opto-Electronics Engineering, Tianjin University. His research interests include laser physics and applications, and cloud/edge computing.