Dynamic Resource Allocation and Computation Offloading for Edge Computing System

Zheng Chang, Liqing Liu, Xijuan Guo, Tao Chen, Tapani Ristaniemi

Dynamic Resource Allocation and Computation Offloading for Edge Computing System[⋆]

Zheng Chang[1], Liqing Liu[2], Xijuan Guo[2], Tao Chen[3], and Tapani Ristaniemi[1]

[1] University of Jyvaskyla, Faculty of Information Technology, P. O. Box 35, FI-40014 Jyvaskyla, Finland
[2] College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[3] VTT Technical Research Centre of Finland, Espoo, Finland zheng.chang@jyu.fi

Abstract. In this work, we propose a dynamic optimization scheme for an edge computing system with multiple users, where the radio and computational resources, and offloading decisions, can be dynamically allocated with the variation of computation demands, radio channels and the computation resources. Specifically, with the objective to minimize the energy consumption of the considered system, we propose a joint computation offloading, radio and computational resource allocation algorithm based on Lyapunov optimization. Through minimizing the derived upper bound of the Lyapunov drift-plus-penalty function, the main problem is divided into several sub-problems at each time slot and are addressed separately. The simulation results demonstrate the effectiveness of the proposed scheme.

Keywords: Edge computing · Dynamic computation offloading · Lyapunov optimization · Resource allocation.

1 Introduction

In mobile cloud computing (MCC), by offloading the computational tasks to the distant cloud for execution, the system performance, e.g., energy consumption and latency, is able to be improved [1]. Among all different types of MCC technologies, fog/edge computing system, emerges as a proximity solution to provide pervasive and distributed computation services for the MDs, and especially for the Internet-of-Things (IoT) applications with stringent requirement of latency and reliability [2]. In the edge computing system, as the computing capability of the edge node (EN) is not comparable to the traditional cloud center and one EN 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 quality of services (QoS). Accordingly, the usage of the radio resources, such as transmit power and frequency spectrum, and the harvested energy should be carefully coordinated and optimized in line with the offloading decisions. In addition, as

[⋆] Supported by organization x.

the radio environment and the demand for computational resources vary in a fast speed, dynamic scheduling and optimization are more preferred compared to static optimization schemes. However, due to the randomness of radio environment, harvested energy and computation demands, realizing the dynamic optimization is challenging. Therefore, in this paper, our aim is to overcome the obstacles and provide dynamic computation offloading and resource allocation schemes for edge computing system with EH devices.

Most of the researches on the offloading problem focus on designing different and effective static schemes for battery-powered MDs, through optimizing the MD's execution decision, radio resource, and/or computational resource [2]- [7]. Considering a edge computing system, the authors of [2] apply queuing theory to investigate the delay, energy consumption, and payment cost (E&D&P) of offloading processes. Based on the theoretical analysis, a multi-objective optimization problem is then formulated to minimize the formulated cost functions by finding the offloading decisions and power allocation for each MD. In [3], the authors explore the tradeoff between delay and energy consumption in the edge-cloud hybrid computing system. The associated workload allocation problem is addressed accordingly. In [4], the authors propose an optimization framework of offloading to optimize the task allocation decision and the computational resource allocation.

In this work, to address the offloading problem in edge computing, we consider different queue models at different edge computing devices to provide thorough analysis on the delay and energy consumption performance. At the MD, a $M/M/1$ queue is considered and at the EN, a $M/G/1$ queue is assumed. With the derived analytical results, we are able to formulate the system cost, which consists of service latency and energy consumption. 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. Due to the stochastic nature of the radio channel, the request arrival and the amount of harvesting energy, 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 can be addressed accordingly.

2 System Model

We consider the system consisting of N single-core MDs, one AP, and one EN. The set of MDs is denoted as $\mathcal{N} = \{1, 2, \cdots, N\}$. Each MD generates a series of homogeneous service requests in order to execute an application. At the MD, a first-in-first-out (FIFO) queue is considered for storing arriving requests, and the radio interface is used for wireless connection. As a single processor is assumed, the process queue at the MD is assumed as a $M/M/1$ queue. The EH capability enables the MD to obtain energy supply from the environment. The harvested energy used for local task execution and data transmission. The AP is responsible for receiving requests from the MD and delivering data to the EN for further processing. The process queue of EN is modelled as a $M/G/1$ queue. The MD offloads (part of) the computation requests to the EN to enjoy a higher level of quality of computation experience. We assume that the time is slotted and the length of each time slot is τ. We denote the time slot set $\mathcal{T}=\{0,1\cdots,t\cdots,T-1\}$.

2.1 Local Execution Model

The computation requests generated by MD i, $i\in\mathcal{N}$ is assumed to follow Poisson process with an average arrival rate $A_i\left(t\right)$ and within $[A_{i,\min},A_{i,\max}]$. Each request is of data size θ_i. Note that "at time slot t" means the requests are generated at time slot "t" but executed at time slot "$t+1$".

For MD i, some of the computation requests may be locally executed and the rest will be offloaded to the EN. It may also happen that when neither of these computation modes is feasible, e.g., when MD has insufficient energy, and some of the computation requests have to be dropped. The decision of MD i at time slot t is modeled as a vector $\mathbf{p}_i\left(t\right)=\left[p_i^M\left(t\right),p_i^F\left(t\right),p_i^D\left(t\right)\right]$, where $p_i^M\left(t\right)+p_i^F\left(t\right)+p_i^D\left(t\right)=1$. $p_i^M\left(t\right)$ represents the portion that the requests are executed locally at time slot t, $p_i^F\left(t\right)$ denotes the portion that the requests are offloaded to the EN, and $p_i^D\left(t\right)$ 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 the MD. Additionally we assume that $l_i^M\left(t\right)$ denotes the normalized workload on the MD i at time slot t, which shows the occupation of CPU. For example, $l_i^M\left(t\right)=0$ indicates at time slot t, the CPU is totally idle. When considering a $M/M/1$ queue with request arrival rate λ and service rate u, the response time is $R=\frac{1/u}{1-\rho}$, where $\rho=\frac{\lambda}{u}$ [8]. Then, the average response time $D_i^M\left(t\right)$ for local execution of MD i at time slot t is expressed as follows:

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

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

$$E_i^M\left(t\right)=k_i{f^3}_i\left(t\right)D_i^M\left(t\right)=\frac{k_i{f^3}_i\left(t\right)}{u_i^M\left(1-l_i^M\left(t\right)\right)-p_i^M\left(t\right)A_i\left(t\right)}. \tag{2}$$

Nevertheless, if some of the requests cannot be executed due to lack of energy, they have to be dropped. We define a cost coefficient μ_i for the task drop, and accordingly the punishment cost for MD i at time slot t can be expressed as follows:

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

2.2 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 $|\mathcal{K}| = K$. The channels are assumed to be independent and identically distributed (i.i.d) block fading during time slots, i.e. the channels remain static within each time slot, but vary among different time slots. Let B 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 $p_{i,\max}$. 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 assigned to MD i at time slot t. Otherwise, $\rho_{i,k}(t) = 0$. Correspondingly, the uplink data rate $r_{i,k}(t)$ of MD i on subcarrier k at time slot t is expressed as follows:

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

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

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

Correspondingly, we can obtain the uplink transmission time $D_i^{up}(t)$, as follows:

$$D_i^{up}(t) = \frac{p_i^F(t) A_i(t) \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t) B\log_2\left(1 + \frac{p_{i,k}(t) h_{i,k}(t)}{N_0 B}\right)}. \tag{6}$$

Then the energy consumption $E_i^{up}(t)$ of the uplink transmission can be given as follows:

$$E_i^{up}(t) = \sum_{k\in\mathcal{K}} \frac{\rho_{i,k}(t) p_{i,k}(t) p_i^F(t) A_i(t) \theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t) B\log_2\left(1 + \frac{p_{i,k}(t) h_{i,k}(t)}{N_0 B}\right)}. \tag{7}$$

2.3 Fog Execution Model

The EN connecting to the AP can process the offloaded requests and execute the computation task. We consider the connection between the EN and AP is fiber-based with large enough bandwidth and the transmission time from the

AP to EN is ignored. We denote the service rate of the EN as u^F. The pending requests of the MDs are pooled together with a total rate $A_{total}(t)$ which also follows the Poisson process. Therefore, $A_{total}(t)$ is given as follows:

$$A_{total}\left(t\right) = \sum_{i \in \mathcal{N}} p_i^F\left(t\right) A_i\left(t\right). \tag{8}$$

We denote the workload of the EN as $l^F(t)$, which presents the occupied percentage of each server and $l^F(t) < 1$. As a $M/G/1$ queue is considered at the EN , the average response time $D^F(t)$ is given as follows [9]:

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

2.4 Energy Harvesting Model

To model the energy harvesting, a successive energy packet arrival model is considered. 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 in each time slot. The harvested energy is stored in the battery and will be available for further actions. We denote the battery energy level of MD i at the beginning of time slot t as $B_i(t)$. In this work, energy consumed for purposes other than local computation and transmission is ignored for simplicity. The energy consumption $E_{i,total}(t)$ of MD i consists of two parts:

$$E_{i,total}\left(t\right) = E_i^M\left(t\right) + E_i^{up}\left(t\right). \tag{10}$$

where $E_i^M\left(t\right)$ is the energy consumption for local processing and $E_i^{up}\left(t\right)$ is energy consumption for delivering the requests. Note that $E_{i,total}(t)$ should be smaller than the battery level, i.e., $E_{i,total}\left(t\right) \leq B_i\left(t\right)$. Thus, the battery level of MD i evolves as follows,

$$B_i\left(t+1\right) = B_i\left(t\right) - E_{i,total}\left(t\right) + e_i\left(t\right). \tag{11}$$

3 Problem Formulation

The execution cost consists of the execution delay and the task dropping punishment cost. The execution delay $D_i(t)$ at time slot t is derived as follows:

$$D_i(t) = p_i^M\left(t\right) D_i^M\left(t\right) + p_i^F\left(t\right)\left(D_i^{up}\left(t\right) + D^F\left(t\right)\right). \tag{12}$$

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

$$EC_i\left(t\right) = D_i\left(t\right) + \alpha_i C_i^D\left(t\right), \tag{13}$$

where α_i is the weight of task dropping cost. The total weighted execution cost of the system at time slot t is denoted as $\mathit{\Gamma}_{total}(t)$, which is given as

$$\mathit{\Gamma}_{total}\left(t\right) \; = \sum_{i\in\mathcal{N}} \omega_i \left[p_i^M\left(t\right)D_i^M\left(t\right) + p_i^F\left(t\right)\left(D_i^{up}\left(t\right) + D^F\left(t\right)\right) + \alpha_i C_i^D\left(t\right)\right], \tag{14}$$

where ω_i is the weight factor, which reflects the relative importance of MD i. Then we derive the average execution cost $\mathit{\Phi}\left(t\right)$ of the edge computing system during T time slots, which is given in (15).

$$\mathit{\Phi}\left(t\right) = \lim_{T\to+\infty}\frac{1}{T}\sum_{t\in\mathcal{T}} WEC_{total}\left(t\right) = \lim_{T\to+\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{N}} \omega_i \left[p_i^M\left(t\right)D_i^M\left(t\right) + p_i^F\left(t\right)\left(D_i^{up}\left(t\right) + D^F\left(t\right)\right) + \alpha_i C_i^D\left(t\right)\right]. \tag{15}$$

We denote the system decision at time slot t as $\mathbf{V}(t) = \left[\mathbf{p}(t), \boldsymbol{\rho}(t), \boldsymbol{p}_{up}(t)\right]$, $\forall t \in \mathcal{T}$, where $\mathbf{p}(t) = [\mathbf{p}_1(t), \cdots, \mathbf{p}_i(t), \cdots \mathbf{p}_N(t)]$ are execution strategies for all the MDs at time slot t and $\mathbf{p}_i(t) = \left[p_i^M(t), p_i^F(t), p_i^D(t)\right]$ 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 matrix for all MDs at time slot t and $\boldsymbol{\rho}_i(t)$= $[\rho_{i,1}(t), \cdots, \rho_{i,k}(t), \cdots, \rho_{i,K}(t)]$ is the subcarrier assignment vector for MD i at time slot t. $\boldsymbol{p}_{up}(t) = [\boldsymbol{p}_1(t), \cdots, \boldsymbol{p}_N(t)]$ is the uplink transmit power matrix for all the MDs at time slot t and $\boldsymbol{p}_i(t)$= $[p_{i,1}(t), \cdots, p_{i,K}(t)]$ is the set of transmit power for MD i. Thus, the problem can be formulated as shown in **P1**, which is

$$\mathbf{P1}: \min_{\boldsymbol{V}(t)} \; \mathit{\Phi}\left(t\right), \tag{16}$$

s.t.

$$p_i^M\left(t\right) + p_i^F\left(t\right) + p_i^D\left(t\right) = 1, 0 \le p_i^M\left(t\right), p_i^F\left(t\right), p_i^D\left(t\right) \le 1; \tag{17a}$$

$$p_i^M\left(t\right)A_i\left(t\right) - u_i^M\left(1 - l_i^M\left(t\right)\right) < 0; \tag{17b}$$

$$\sum_{i\in\mathcal{N}} p_i^F\left(t\right)A_i\left(t\right) - u^F\left(1 - l^F\left(t\right)\right) < 0; \tag{17c}$$

$$0 < p_{i,k}\left(t\right) < p_{i,\max}; \tag{17d}$$

$$\sum_{i\in\mathcal{N}} \rho_{i,k}\left(t\right) \le 1, \quad \rho_{i,k} \in \left\{0, 1\right\}; \tag{17e}$$

$$E_{i,total}\left(t\right) \le B_i\left(t\right); \tag{17f}$$

$$i \in \mathcal{N}, t \in \mathcal{T}, k \in \mathcal{K}. \tag{17g}$$

As we can see, the MDs' decisions are coupled among different time slots due to the constraints (17f), which makes the problem difficult to be tackled. As presented in [5], 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, the system operation can be optimized by ignoring (17f). Thus, the problem can be modified as follows:

$$\begin{aligned} &\mathbf{P1} : \min_{\boldsymbol{V}(t)} \ \mathit{\Phi}(t) \\ &(17\mathrm{a}) - (17\mathrm{e}), (17\mathrm{g}) \end{aligned} \tag{18}$$

$$E_{i,total}(t) \in \left[E_i^{\min}(t), E_i^{\max}(t)\right] \tag{19}$$

For simplify, we consider $E_i^{\min}(t) = 0$. For $\mathbf{P1}$, a stochastic optimization problem is formulated with decision variables of the execution strategy, the uplink transmit power and the subcarrier assignment. By addressing the deterministic per-time slot problem, we can obtain the total optimal decisions in a stochastic manner.

4 Proposed Solution

Lyapunov optimization is an efficient framework for designing online control algorithm without requiring any prior knowledge [5]. In order to present the proposed solution, we firstly define the Lyapunov function as follows:

$$L(\boldsymbol{B}(t)) = \frac{1}{2}\sum_{i\in\mathcal{N}} {B_i}^2(t), \tag{20}$$

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

$$\mathit{\Delta}(\boldsymbol{B}(t)) {=} E\left[L(\boldsymbol{B}(t+1)) - L(\boldsymbol{B}(t))\left|\boldsymbol{B}(t)\right.\right]. \tag{21}$$

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

$$\mathit{\Delta}_V(\boldsymbol{B}(t)) = \mathit{\Delta}(\boldsymbol{B}(t)) + V\mathcal{E}\left[\mathit{\Gamma}_{total}(t)\left|\boldsymbol{B}(t)\right.\right], \tag{22}$$

where $V \in (0, +\infty)$ is a control parameter. Then we will find an upper bound of $\mathit{\Delta}(\boldsymbol{B}(t))$ 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 (18) and (19), the Lyapunov drift-plus-penalty function $\mathit{\Delta}_V(\boldsymbol{B}(t))$ is upper bounded, i.e.,

$$\begin{aligned} \mathit{\Delta}_V(\boldsymbol{B}(t)) \le \kappa + &\sum_{i\in\mathcal{N}} \left\{B_i(t)\left[e_i(t) - E_{i,total}(t)\right]\right\} \\ &+ VE\left[\mathit{\Gamma}_{total}(t)\left|\boldsymbol{B}(t)\right.\right], \end{aligned} \tag{23}$$

where κ is a constant, which is denoted as

```
Algorithm 1 Proposed online algorithm
Step 1: at the beginning of the time slot $t$, obtain $\boldsymbol{B}(t)$.
Step 2: through solving the problem P2, determine the system decision set $\boldsymbol{V}(t) = [\boldsymbol{p}(t), \boldsymbol{\rho}(t), \boldsymbol{p}_{up}(t)]$, to minimize the P2.
    $\min_{\boldsymbol{V}(t)} \sum_{i\in\mathcal{N}} \{B_i(t)[e_i(t) - E_{i,total}(t)]\} + V\mathcal{E}[\Gamma_{total}(t)|\boldsymbol{B}(t)|]$
s.t. (18), (19)
Step 3: set $t = t+1$, update $\boldsymbol{B}(t)$ , repeat Step 1 and Step 2, until obtain the system decisions of all the time slots.
```

$$\kappa = \sum_{i\in N}\left[\frac{(e_i^{\max}(t))^2 + (E_i^{\max}(t))^2}{2}\right]. \tag{24}$$

Due to the space limitation, we omit the proof here. The key idea of the proposed algorithm is to minimize the upper bound of $\Delta_V(\boldsymbol{B}(t))$ in the right-hand side of (23). The proposed algorithm is displayed in Algorithm 1.

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

4.1 Optimal Execution Strategy

Firstly, we seek the optimal execution 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 denoted as follows:

$$\min_{\boldsymbol{p}(t)} \sum_{i\in\mathcal{N}} -B_i(t)E_{i,total}(t) + V\sum_{i\in\mathcal{N}} \omega_i\left[D_i(t) + \alpha_i C_i^D(t)\right] \tag{25}$$

s.t.

$$(17\mathrm{a}) - (17\mathrm{g}), (17\mathrm{g}), (19)$$

It can be found that (17c) is a coupled constraint, which includes various decision variables of different MDs. Similarly to the ones in [6], 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.

4.2 Optimal Power Allocation and Subcarrier Assignment

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:

$$\min_{\{\boldsymbol{\rho}(t),\boldsymbol{p}(t)\}} \sum_{i\in\mathcal{N}} -B_i(t)E_i^{up}(t) + V\sum_{i\in\mathcal{N}} \omega_i p_i^F(t)D_i^{up}(t), \tag{26}$$

s.t.

$$0 < p_{i,k}(t) < p_{i,\max}, \tag{27a}$$

$$\sum_{i\in\mathcal{N}} \rho_{i,k}(t) \le 1, \quad \rho_{i,k} \in \{0,1\}, \tag{27b}$$

$$E_i^{up}(t) < E_i^{\max}(t), \tag{27c}$$

$$i \in \mathcal{N}, k \in \mathcal{K}. \tag{27d}$$

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 (27). We can find that the **SP2'** is a mixed-integer programming problem, which involves the joint optimization of both continuous variables $p_{i,k}(t)$ and integer variables $\rho_{i,k}(t)$. Next, we will propose an algorithm to solve the problem. Firstly, we introduce an average offloading priority function [7], and it is defined as follows:

$$\begin{aligned}\textbf{SP2'}: \min_{\{\boldsymbol{\rho}(t),\boldsymbol{p}(t)\}} \quad & \sum_{i\in\mathcal{N}} -B_i(t) \sum_{k\in\mathcal{K}} \frac{\rho_{i,k}(t)\, p_{i,k}(t)\, p_i^F(t)\, A_i(t)\,\theta_i \tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, B\log_2\left(1+\frac{p_{i,k}(t)h_{i,k}(t)}{N_0 B}\right)} \\ & + V\sum_{i\in\mathcal{N}} \omega_i p_i^F(t) \left(\frac{p_i^F(t)\, A_i(t)\,\theta_i\tau}{\sum\limits_{k\in\mathcal{K}} \rho_{i,k}(t)\, B\log_2\left(1+\frac{p_{i,k}(t)h_{i,k}(t)}{N_0 B}\right)} \right)\end{aligned} \tag{28}$$

$$\begin{aligned}&\psi_{i,k,t}(\omega_i,\tau,h_{i,k}(t)) \\ &= \begin{cases} \frac{\omega_i N_0 B}{h_{i,k}(t)} \left[v_i(t)\ln v_i(t) - v_i(t) + 1\right], & v_i(t) \ge 1, \\ 0, & v_i(t) < 1, \end{cases}\end{aligned} \tag{29}$$

where the constant $v_i(t)$ is defined as $v_i(t) = \frac{Bh_{i,k}(t)\tau c_0}{N_0 \ln 2}$ and c_0 is a predefined 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 $\tilde{h}_i(t) = \frac{\sum\limits_{k\in\mathcal{K}_2(t)} h_{i,k}(t)}{|\mathcal{K}_2(t)|}$, 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 displayed in Algorithm 2.

In the proposed algorithm, we need to find the optimal power allocation, which involves addressing the following **SP2"**, which is

Algorithm 2 Subcarrier allocation algorithm

```
Input:
  At beginning of time slot $t$, obtain $\Psi(t)$, $h_{i,k}(t)$,$\mathcal{K}_1(t)$, $\mathcal{K}_2(t)$, and $\tilde{h}_i(t)$;
Obtain the total integer number of subcarriers:
  Solving the optimal solution $\{n_i^*(t), \tilde{p}_i^*(t)\}$ of the SP2";
Subcarrier allocation:
while $\tilde{\mathcal{N}} \neq \varnothing$, do
    (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 $\tilde{\mathcal{N}} = \tilde{\mathcal{N}} \setminus \{i'\}$;
end while
Transmit power allocation
  Solving the optimal solution of SP2"'.
return $\{\rho_{i,k}^*(t), p_{i,k}^*(t)\}$
```

$$\mathbf{SP2"}: \min_{\{\boldsymbol{n}_i(t), \tilde{\boldsymbol{p}}_i(t)\}} \sum_{i \in \mathcal{N}} \Bigg\{ \frac{-B_i(t)\,\tilde{p}_i(t)\,p_i^F(t)\,A_i(t)\,\theta_i \tau}{B\log_2\left(1 + \frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 B}\right)} + \frac{V\omega_i\left[p_i^F(t)\right]^2 A_i(t)\,\theta_i\tau}{n_i(t)\,B\log_2\left(1 + \frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 B}\right)} \Bigg\}, \tag{30}$$

s.t.

$$\sum_{i \in \mathcal{N}} n_i(t) \leq |\mathcal{K}_2(t)|, \tag{31a}$$

$$\tilde{p}_i(t) \leq p_{i,\max}, \tag{31b}$$

$$\frac{\tilde{p}_i(t)\,p_i^F(t)\,A_i(t)\,\theta_i\tau}{B\log_2\left(1 + \frac{\tilde{p}_i(t)\tilde{h}_i(t)}{N_0 B}\right)} \leq E_i^{\max}(t), \tag{31c}$$

where $n_i(t)$ is the total integer number of subcarriers that allocated to MD i at time slot t. We can also find that **SP2"** is a mixed integer programming including a coupled constraint (31a). Thus, we can address it with semi-smooth Newton method, which is similar with [6]. Then with 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 with the highest offloading priority principle. After searching for the highest offloading priority $\psi_{i',k',t}$ over 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

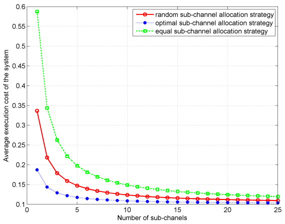

Fig. 1. The effect of subcarrier allocation

at time slot t over the assigned subcarriers $\mathcal{Z}_i(t)$ is obtained by minimizing the problem **SP2"'**

$$\begin{aligned}\mathbf{SP2"'}: \min_{p_{i,k'}(t), k' \in \mathcal{Z}_i(t)} & \sum_{k' \in \mathcal{Z}_i(t)} \frac{-B_i(t)\, p_{i,k'}(t)\, p_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k' \in \mathcal{Z}_i(t)} B \log_2\left(1 + \frac{p_{i,k'}(t) h_{i,k'}(t)}{N_0 B}\right)} \\ & + \frac{V \omega_i \left(p_i^F(t)\right)^2 A_i(t)\, \theta_i \tau}{\sum\limits_{k' \in \mathcal{Z}_i(t)} B \log_2\left(1 + \frac{p_{i,k'}(t) h_{i,k'}(t)}{N_0 B}\right)},\end{aligned} \tag{32}$$

s.t.

$$0 < p_{i,k'}(t) \le p_{i,\max}, k' \in \mathcal{Z}_i(t), \tag{33a}$$

$$\sum_{k' \in \mathcal{Z}_i(t)} \frac{p_{i,k'}(t)\, p_i^F(t)\, A_i(t)\, \theta_i \tau}{\sum\limits_{k' \in \mathcal{Z}_i(t)} B \log_2\left(1 + \frac{p_{i,k'}(t) h_{i,k'}(t)}{N_0 B}\right)} \le E_i^{\max}(t), \tag{33b}$$

We can see that the formulated problem **SP2"'** is similar with the problem investigated in [2]. Then, we can solve it with Interior Point Method (IPM), the details of which can be found in [2].

5 Performance Evaluations

In this section, extensive simulations are conducted to illustrate the effectiveness of the proposed algorithm. The simulation parameters are similar to the one used in [2] and [6]. First, we illustrate the relationship of the average execution cost of the system versus the number of subcarriers with 6 MDs in Fig. 1. It can be observed that with the optimal subcarrier allocation strategy, the average execution cost of the system is the smallest among all three schemes. Moreover, as shown in this figure, with the increasing of the number of subcarriers, the average execution cost becomes smaller, as the MDs have sufficient choices to offload the requests to the EN to reduce the execution delay. In this way, the dropped requests would also be reduced.

Then we show the total execution cost of the system versus the number of MDs in the system when the number of subcarriers is fixed in Fig. 2. It can be

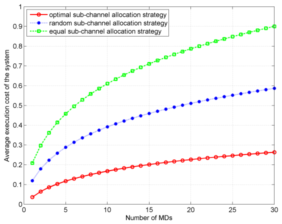

Fig. 2. The effect of the number of MDs

observed that the average execution costs are increasing when the number of MDs increases, which means that the execution delay or the punishment cost become larger 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 delay can be induced. Thus, the MDs have to execute more requests locally or drop them, which leads to a larger execution cost.

6 conclusion

In this paper, we propose a dynamic optimization scheme for an edge computing system with multiple users, where the radio and computational resources, and offloading decisions, can be dynamically allocated with the variation of computation demands, radio channels and the computation resources. Specifically, with the objective to minimize the energy consumption of the considered system, we propose a joint computation offloading, radio and computational resource allocation algorithm based on Lyapunov optimization. Through minimizing the derived upper bound of the Lyapunov drift-plus-penalty function, the main problem is divided into several sub-problems at each time slot and are addressed separately. The simulation results demonstrate the effectiveness of the proposed scheme.

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. 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.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. A. Lazar, "The throughput time delay function of an M/M/1 queue (Corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 914-918, Nov. 1983.
9. R. E. Machol, "Queue theory," *IRE Transactions on Education*, vol. E-5, no. 2, pp. 99-105, Nov. 2007.