SCADS: Simultaneous Computing and Distribution Strategy for Task Offloading in Mobile-Edge Computing System

Conference Paper · July 2019

DOI: 10.1109/ICCT.2018.8599958

Anderson M.H. Jiao
Beijing University of Posts and Telecommunications

SCADS: Simultaneous Computing and Distribution Strategy for Task Offloading in Mobile-Edge Computing System

Haoran Liu, Haoyue Zheng, Minghan Jiao, Guoxuan Chi
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: {liu_haoran, chiguoxuan12345}@bupt.edu.cn

***Abstract*—Mobile edge computing (MEC) has emerged as a prominent technique to improve the quality of computation experience for mobile devices in the fifth-generation (5G) networks. However, the design of computation task scheduling policies for MEC systems inevitably encounters a challenging latency optimization problem. Due to the limited radio and computational resources in communication system, a more efficient latency-optimal scheduling policy is urgently needed to meet the ever-increasing computation demands of many new mobile applications. In this paper, we formulate an optimization problem based on partial offloading strategy and transform it into a piecewise convex problem, getting the latency-optimal point by means of sub-gradient method. A simplified algorithm is further put forward to achieve close-to-optimal performance in polynomial time. Therefore, we conclude a simultaneous computing and distribution strategy called SCADS. Simulation results are provided to demonstrate the advantages of our proposed algorithms compared with other baseline strategies.**

Keywords-obile-edge computing (MEC); simultaneous computing and distribution; task offloading; latency-optimal

I. Introduction

With the popularity of mobile devices, computation-intensive mobile applications such as online VR games, high-definition video conferences and 3D modeling are becoming more prevalent [1]-[3]. However, most of these mobile devices possess limited resources (i.e., limited battery energy and computation capability of local CPUs), and thus may suffer from unsatisfactory computation experience. The emergency of Cloud Computing (CC) [4] and Mobile Cloud Computing (MCC) [5] helped to reduce the workload of mobile devices through offloading tasks to cloud servers. Nevertheless, servers' congestion usually results in high latency so that the quality of experience (QoE) [6] for end-users is seriously weakened. As a promising paradigm to resolve this kind of issue, mobile-edge computing (MEC) has emerged in recent years. MEC offers computation capability within the radio access network (RAN) [7]-[9]. By offloading computation-intensive tasks from the mobile devices to the nearby MEC servers, the quality of computation experience, including the latency and device energy consumption, could be greatly improved, which is extremely essential in 5G wireless networks.

Computation offloading strategies in MEC systems determine the efficiency and computation performance of the systems. Many researchers focused on single-user, multi-user and heterogeneous-user MEC systems. In [10], the author proposed a heuristic load-balancing program-partitioning algorithm to address the concurrent tasks for computation-intensive applications. Using a semi-Markovian decision process based control method, offloading tasks will transfer from mobile devices to the cloud more effectively [11]. The author of [12] optimized the allocation of communication and computation resources in multi-user mobile cloud computing jointly. Additionally, a stochastic task arrival model based on the Lyapunov optimization [13] has been put forward to solve energy-latency tradeoff problem for multi-user MEC systems. An energy-efficient offloading strategy for MEC systems based on binary offloading in 5G networks has been proposed in [14]. Most of these works focused on binary offloading and optimization of energy consumption. However, a latency-optimal strategy based on partial offloading is also an urgent need of the near-future 5G networks [15]. Therefore, we put forward a latency-optimal offloading strategy called SCADS, which will be furtherly analyzed.

In SCADS, a multi-user MEC system with multiple independent tasks will be discussed. The system has to allocate all the resources efficiently under the radio and computational resource constraints so that the optimization of task offloading proportion and resource allocation are both needed for MEC systems. However, keeping balance of resource allocation and task segmentation is really difficult, which makes latency-optimal partial offloading can hardly be achieved by conventional convex optimization methods. Therefore, it's challenging for us to achieve the lowest latency. After analyzing the objective functions, we transform it into a piecewise convex problem, which can be easily solved by using sub-gradient approach. Meanwhile, a polynomial-time-complexity algorithm based on a common scenario is further proposed to simplify the strategy.

The rest of this paper is organized as follows. We introduce the system model in Section II. The latency optimization problem based on resource allocation and data segmentation is analyzed and formulated in Section III. In Section IV, the solutions to the problems are shown, while the optimal and close-to-optimal algorithms in SCADS are put forward. Simulation results are shown in Section V and we will conclude this paper in Section VI.

978-1-5386-7635-6/18/$31.00 ©2018 IEEE

II. SYSTEM MODEL

Figure 1 shows MEC system for multi-users in 5G networks. In this figure, there are several Base Stations (BS) cover various mobile devices where many computation-intensive applications are working. Edge Servers (ES) that under BS's control are capable of computing. Re-arranging the computation offloading strategy at intervals will happen in the system. The system will also transmit different data segments to ES and local mobile devices for parallel computing. Finally, user's mobile devices will aggregate all processed data.

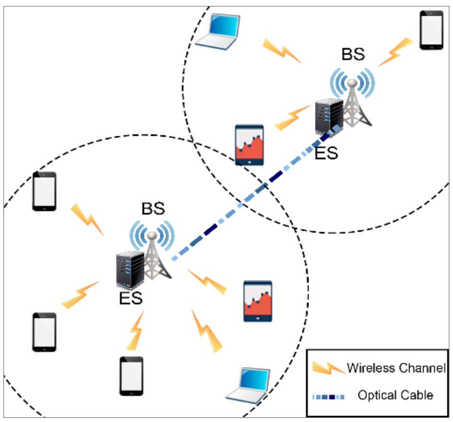

Figure 1. MEC system in 5G networks

A. Task Model

Assume that an MEC system includes N mobile devices. Denote the task set as $T = \{T_1, T_2, \dots, T_N\}$, where T_d represents a task requested by device d. Each task is characterized by a three-tuple of parameters $\langle S_d, W_d, \gamma_d \rangle$, where S_d (in bits) and W_d (in CPU cycles/bit) denote the size and workload of the task respectively, while $\gamma_d \in (0, +\infty)$ indicates the data growth ratio (i.e., one-bit raw data will be processed into γ_d bits). With these parameters, SCADS can be established and offer a better quality of computation experience for users.

B. Computation Model

Suppose there are K BSs in the system, which denoted by a set $K = \{1, 2, \dots, K\}$. V_d and V_k (both in CPU cycles/second) describe the computing capacity of device d and ES k respectively. In the practical multi-user MEC system, ESs are often used by numerous devices at the same time. Therefore, in order to analyze one task's computing process more specifically, let V_k^d denotes the computing capacity of ES k allocated to device d. Suppose that MEC servers are equipped with multi-core CPUs for computing and can be also virtualized into several virtual machines, which makes the concurrent of multi-tasks feasible.

C. Communication Model

Consider the MEC system is operating in the OFDMA mode. Under such circumstances, all BSs are working in the same frequency band with a bandwidth BW. The spectrum is divided into M orthogonal sub-channels which denoted as a set $M = \{1, 2, \dots, M\}$. The bandwidth of each sub-channel is identical. Denote $\alpha_m^d = \{0, 1\}$ as a state indicator of sub-channel m. For example, if sub-channel m is allocated to device d, then $\alpha_m^d = 1$, otherwise $\alpha_m^d = 0$. Therefore the data transmission rate can be expressed as

$$r_d = \sum_{m=1}^{M} \alpha_m^d \cdot \frac{BW}{M} \log_2\left(1 + \frac{P_m \cdot |g_m^d|^2}{N_0}\right) \tag{1}$$

where P_m and N_0 denote the transmission power of BS in channel m and the variance of Additive White Gaussian Noise respectively, while g_m^d indicates the channel gain of device d on channel m . Since g_m^d is a random variable, data transmission rate r_d is also a random variable. Expected value used to evaluate the transmission rate r_d can be achieved according to [16]. To facilitate the understanding, we define the channel occupancy of device d as $\rho_d = \frac{1}{M} \cdot \sum_{m=1}^{M} \alpha_m^d$.

D. Optimization Model

Resource allocation and task segmentation strategies will be both executed for device d in the system. Denote $\theta_d \epsilon [0,1]$ as the segmentation proportions and define the optimization strategy executed at BS as well as the detailed procedure can be described like the following procedure

The Internet Content Providers (ICP) will send the raw data to Base Stations (BS) for computing. The size of the task is S_d bits, while $\theta_d S_d$ bits data will be processed on ES and the remaining $(1 - \theta_d)S_d$ bits raw data will be transmitted to device d for the rest computation. After ES complete its computation, $\gamma_d \theta_d S_d$ bits processed data will be transmitted to devices. Detailed sub-processes are shown in Figure 2.

TABLE I. DELAY EXPRESSIONS

No.	Latency Expressions	Sub-processes
①	$D_c^c = \sum_{d=1}^{n}\left(\frac{W_d S_d}{V_c^e}\right)\theta_d$	Process raw data on ES
②	$D_c^t = \sum_{d=1}^{n}\left(\frac{\gamma_d S_d}{R_e}\right)\theta_d$	Transmit processed data to UE
③	$D_d^t = \frac{S_d}{\rho_d R_d}(1 - \theta_d)$	Transmit raw data to UE
④	$D_d^c = \frac{W_d S_d}{V_d}(1 - \theta_d)$	Process raw data on UE

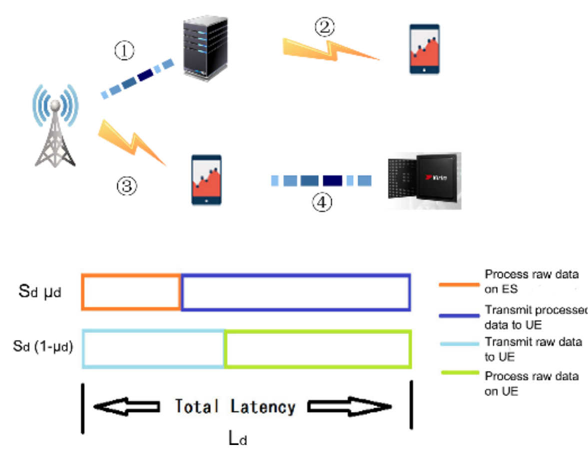

Figure 2. SCADS in MEC System

III. Problem Formulation

The optimization problem can be formulated by analyzing the concurrency relations of different sub-processes mentioned in Section II. So the delay expression of device d can be shown as

$$D_d = \begin{cases} max\{D_d^t + D_d^c\ ,\ D_c^t + D_c^c\}, & D_c^c \geq D_d^t, \\ D_d^t + max\{D_c^t\ ,\ D_d^c\}, & D_d^t > D_c^c. \end{cases} \quad (2)$$

This strategy is intended to minimize the weighted-sum delay under radio and computational resource constraints, which indicates that two optimization problems can be formulated in the same form. Denote a positive weight factor set $\{\omega_n\}$ that satisfies $\sum_{d=1}^{N} \omega_d = 1$. Accordingly, the optimization problem is shown as

$$min\ \sum_{d=1}^{N} \omega_d\, D_d \quad (3a)$$

$$\boldsymbol{s.t.}\ \sum_{d=1}^{N} \rho_d \leq 1, \quad \rho_d > 0 \quad (3b)$$

$$\sum_{d=1}^{N} V_k^d \leq V_k, \quad V_k^d > 0 \quad (3c)$$

where (3b) and (3c) are the radio and computational resource constraints respectively. By relaxation and rounding method, the discrete variable ρ_n can be converted into a continuous variable for ease of notion.

An optimization problem is formulated and the weighted-sum delay can be minimized through solving the problem.

IV. Problem Solving

According to the analysis of Section III, the optimal and close-to-optimal strategies are set up in the following section.

A. Problem Transformation

The expression of D is complicated with three variables coupled with one another, so the optimization problem (3a) can hardly be solved directly. Suppose a special case

$$\begin{cases} D_c^c = D_d^t, \\ D_c^t = D_d^c, \end{cases} \quad (4)$$

where indicates that the computing capability of device d equals to its transmission capacity (i.e., $\mathrm{R}_d\rho_d W_d = \sqrt{\gamma_d V_k^d V_d}$). Thus, two special segmentation proportions are solved as

$$\begin{cases} \theta_d^1 = \frac{V_k^d}{V_k^d + \mathrm{R}_d\rho_d W_d}, \\ \theta_d^2 = \frac{\mathrm{R}_d\rho_d W_d}{\gamma_d V_d + \mathrm{R}_d\rho_d W_d}. \end{cases} \quad (5)$$

The optimization problem can be discussed in two cases. When $\theta_d^1 \leq \theta_d^2$, the minimum value can be achieved when $D_d^t + D_d^c = D_c^t + D_c^c$. If $\theta_d = \theta_d^2$, we can get the minimum value in the case of $\theta_d^1 \geq \theta_d^2$. Therefore, the data segmentation proportion is shown as

$$\theta_d = \begin{cases} \frac{V_k^d(V_d + R_d\rho_d W_d)}{(1+\gamma_d)V_d V_k^d + R_d\rho_d W_d(V_d + V_k^d)}, & \theta_d^1 \leq \theta_d^2, \\ \frac{R_d\rho_d W_d}{\gamma_d V_d + R_d\rho_d W_d}, & \theta_d^1 > \theta_d^2. \end{cases} \quad (6)$$

Substitute (6) into (2), the delay expression is rewritten as

$$\widehat{D_d} = \begin{cases} \frac{\gamma_d V_k^d(V_d + R_d\rho_d W_d) + V_d(R_d\rho_d W_d)^2 S_d}{(1+\gamma_d)V_d V_k^d + (R_d\rho_d W_d)^2(V_d + V_k^d)}, & \rho_d \geq \frac{\sqrt{\gamma_d V_k^d V_d}}{R_d W_d}, \\ \frac{\gamma_d S_d}{\gamma_d V_d + R_d\rho_d W_d}, & \rho_d < \frac{\sqrt{\gamma_d V_k^d V_d}}{R_d W_d}. \end{cases} \quad (7)$$

B. Optimal Strategy

The transformed delay expression (7) is continuous but non-differential at piecewise point, therefore, classical KKT conditions can not be applied. Substitute (7) into (3a) and then we are able to solve the optimization problem by the sub-gradient method. Sub-gradient function is shown as

$$\boldsymbol{g} = \begin{cases} \boldsymbol{\partial}\left(\sum_{d=1}^{N} V_k^d\right) \\ \boldsymbol{\partial}\left(\sum_{d=1}^{N} \rho_d\right) \\ \boldsymbol{\partial}\left(\sum_{d=1}^{N} \omega_d \widehat{D_d}\right) \end{cases} \quad (8)$$

where $\boldsymbol{\partial}\left(\sum_{d=1}^{n} \omega_d \widehat{D_d}\right)$is a piecewise expression. Referring to non-differential convex optimization theory, the problem can be solved by iteration expression shown as

$$x^{(n+1)} = x^{(n)} + \phi_n g^{(n)} \quad (9)$$

where $\boldsymbol{x}$ denotes the resource allocation strategy. Procedure of sub-gradient algorithm is presented in Algorithm 1.

Algorithm 1 Sub-gradient algorithm

1: **Input** $V_d = [V_1, \dots, V_N]$, $V_{ES} = [V_{S1}, \dots, V_{SK}]$, $P_{BS} = P_0$, $T_d = [T_1, \dots, T_N]$, $R_d = [R_1, \dots, R_N]$, $\epsilon = \epsilon_0$ and $n = 0$.
2: **Output** $\boldsymbol{x}^{(n)}$ and $\boldsymbol{\theta} = [\theta_1, \dots, \theta_n]$.
3: **Do**
4: Set $F_{(old)} = F_{(new)}$.
5: Update $\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} + \phi_n \boldsymbol{g}^{(n)}$.
6: Update $F_{(new)} = \sum_{d=1}^{n} w_d \hat{L}_d$.
7: Update $\boldsymbol{\theta}$ according to (7).
8: Update $\boldsymbol{g}^{(n)}$ according to (9).
9: Set $n = n + 1$.
10: **While** $F_{(old)} - F_{(new)} \geq \epsilon_0$.

While it is an algorithm with polynomial time complexity $O(\frac{1}{{\epsilon_0}^2})$, ϵ_0 has to be an infinitesimal value to achieve the optimal latency causing a tremendous execution delay, which is completely intolerable in practical scenarios.

C. Close-to-optimal Strategy

The resource allocation and data segmentation are dynamic in practical scenarios because of the change of mobile devices' locations and requested contents. The proposed algorithm needs to be executed every few seconds. So we consider a common case so as to simplify Algorithm 1.

Since computation-intensive applications usually take a long execution time, some of the new features in 5G (e.g. higher bandwidth, massive MIMO and beamforming technology) guarantee that the data transmission rate is far greater than processing rate. So $\theta_d^1 \leq \theta_d^2$ is always true. Also, the equation $D_d^d + D_d^c = D_c^d + D_c^c$ can also be approximatively written as $D_d^c = D_c^t + D_c^d$ so that a new approximate segmentation proportion can be calculated. Substitute the new proportion into (3) and $\overline{D}_d$ can be shown as

$$\overline{D_d} = \frac{(\rho_d S_d W_d+\gamma_d V_c^d)}{\gamma_d V_d V_c^d+\rho_d S_d W_d(V_d+V_c^d)} \tag{10}$$

which is a differential function. Thus, the convex problem can be solved by classic KKT condition. Construct a Lagrange function of $\sum_{d=1}^{n}\omega_d \overline{D}_d$ shown as

$$F_d = \sum_{d=1}^{N}\omega_d \overline{D}_d + \delta\left(\sum_{d=1}^{N}\rho_d - 1\right) + \varepsilon\left(\sum_{d=1}^{N}V_c^d - V_c\right) \tag{11a}$$

$$\boldsymbol{s.t.}\ \sum_{d=1}^{N}\rho_d \le 1, \rho_d > 0, \delta\left(\sum_{d=1}^{N}\rho_d - 1\right) = 0 \tag{11b}$$

$$\sum_{d=1}^{N}V_c^d = V_c, V_c^d \ge 0, \varepsilon\left(\sum_{d=1}^{N}V_c^d - V_c\right) = 0 \tag{11c}$$

where $\delta \ge 0$ and $\varepsilon \ge 0$ are Lagrange multipliers associated with radio and computational resource constraints respectively. By computing the partial derivatives of (11a) and making them equal to zero, the close-to-optimal resource allocation strategy can be expressed as

$$\begin{cases} \rho_d = \dfrac{\left(\sqrt{\frac{\gamma_d \sum \omega_d R_d}{\delta}}-\gamma_d V_d\right)V_k^d}{R_d W_d (V_d+V_k^d)} \\ V_k^d = \dfrac{\left(\sqrt{\frac{\gamma_d}{\varepsilon}}-V_d\right)R_d \rho_d W_d}{\gamma_d V_d+R_d \rho_d W_d} \end{cases} \tag{12}$$

After transforming (11b) and (11c) into equality constraints, we get the optimal resource allocation by solving a linear system of equations. Substitute (12) into (6), the offloading proportion θ_d can be solved.

The procedure of the strategy is showed in Algorithm 2.

Algorithm 2 Close-to–optimal offloading algorithm

1: **Step 1**
2: Update V_d, T_d by using context-awareness
3: Update R_d by using context-awareness
4: **Step 2**
5: Calculate V_k^d and ρ_d according to (12).
6: Calculate θ_d according to (6).
7: Operate data segmentation and execution.

Since explicit expressions have been given in (6) and (12), the time complexity of Algorithm 2 is $O(\mathrm{n})$, which determined by the number of mobile devices.

V. NUMERICAL RESULTS

To verify our proposed two strategies, we first compare weighted-sum delay between different strategies and then evaluate resource allocation between mobile devices with different computing capacities. We perform 500 independent repeated trails in order to reduce the randomness. Most simulation parameters are listed in Table II.

TABLE II. MOST SIMULATION PARAMETERS

Parameters	Values
Cellular radius	100 m
BS power	34 dBm
Noise power density	-174 dB/Hz
Task size	1-10 Mbits
Data growth ratio	3.0
Bandwidth	200 MHz
Task workload	100-1000 cycles/s
ES capacity	10^{11} cycles/s

A. *Weighted-sum System Latency*

In this section, simulations on weighted-sum system delay of five models are demonstrated. Computing capacities of all devices follow the uniform distribution with $V_d \in [0.5 \times 10^9, 4.5 \times 10^9]$. The locations of mobile devices follow the two-dimension uniform distribution. Simulations on system delay versus the number of devices and task workload of five different models are shown in Figure 3.

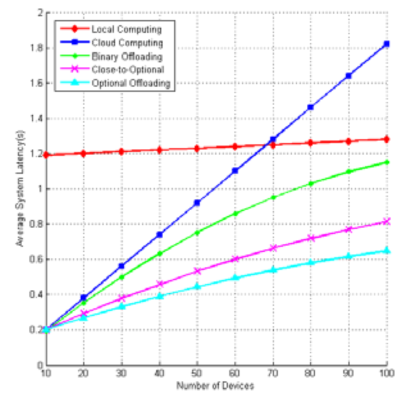

Figure 3. Average Delay vs. Number of Devices

Figure 3 shows the delay of five models all increase with the number of devices since computational and radio resources are limited. The derivative of local computing model is the smallest since the radio resource is relatively adequate. With the increment of mobile devices, the delay of cloud computing model becomes longer than the local computing model since computing capacity of MEC server is limited. If the number of devices is high enough, computing capacity allocated to each device would be much smaller than the computing capacity of local device. Two partial offloading models perform better than binary offloading model since the former one processes in parallel. Partial offloading models reduce 17.6% to 43.4% delay compared with binary offloading.

As shown in Figure 4, with the increment of task workload, system latency becomes longer, performance gaps between our strategies and other strategies become more evident. Compared with cloud computing, partial offloading reduce about 58% delay and thereby significantly improve QoE for users who require computation-intensive services.

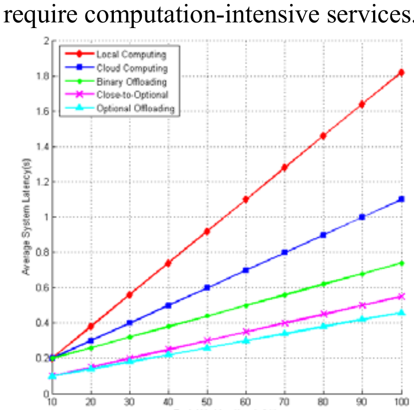

Figure 4. Average Delay vs. Task Workload

B. Resource Allocation

In this subsection, we analyze the radio and computational resource allocation of tasks requested by five mobile devices with different computing capacities. Note that *Device 3* is a baseline with a task size $S_3 = 5$ Mbits.

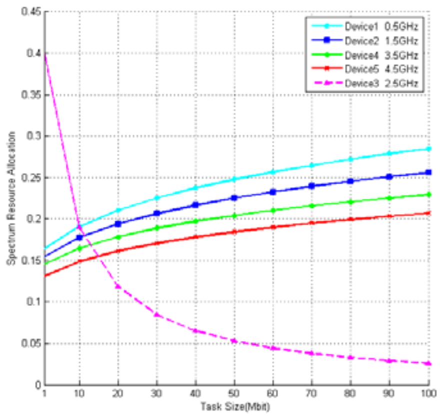

Figure 5. Spectrum Resource Allocation

Figure 5 indicates that more resources will be allocated to devices with heavier tasks. Moreover, the system prefers to allocate resources to devices with inadequate capacities.

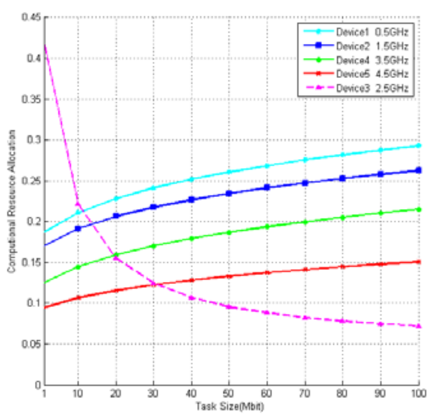

Figure 6. Computatation Resource Allocacation

According to Figure 6, the proportion gaps between five devices are much larger than that in Figure 5. because the limitation of computational resource is the major problem for computation-intensive services.

VI. Conclusion

In this paper, the delay optimization of MEC systems in 5G networks was investigated and a simultaneous computing and distribution strategy (SCADS) for task offloading was proposed. We formulated the optimization problem based on partial offloading. After being transformed into a piecewise convex problem, the problem can be solved by sub-gradient method. To reduce the complexity of the strategy, a simplified algorithm based on a common scenario was derived, which has a great potential for practical implementation. The effectiveness of our proposed strategy is verified by numerical results of our experiments.

References

[1] Y. Mao, J. Zhang and K. B. Letaief, "Joint Task Offloading Scheduling and Transmit Power Allocation for Mobile-Edge Computing Systems," 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, 2017, pp. 1-6.

[2] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," in IEEE Communications Surveys & Tutorials, vol. PP, no. 99, pp. 1-1.

[3] J. Liu et al., "Delay-optimal computation task scheduling for mobile-edge computing systems," 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, 2016, pp. 1451-1455.

[4] G. Orsini, D. Bade and W. Lamersdorf, "Computing at the Mobile Edge: Designing Elastic Android Applications for Computation Offloading," 2015 8th IFIP Wireless and Mobile Networking Conference (WMNC), Munich, 2015, pp. 112-119.

[5] X. Chen et al., "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," in IEEE/ACM Transactions on Networking, vol. 24, no. 5, pp. 2795-2808, October 2016.

[6] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," in IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628-1656, thirdquarter 2017.

[7] K. Zhang et al., "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," in IEEE Access, vol. 4, pp. 5896-5907, 2016.

[8] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016, pp. 1-8.

[9] S. Sardellitti, G. Scutari and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," in IEEE Transactions on Signal and Information Processing over Networks, vol. 1, no. 2, pp. 89-103, June 2015.

[10] M. Jia, J. Cao, and L. Yang, “Heuristic offloading of concurrent tasks for computation-intensive applications in mobile cloud computing,” in Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM WKSHPS), Toronto, Canada, Apr. 2014, pp. 352–357.

[11] S. Chen, Y. Wang, and M. Pedram, “A semi-Markovian decision process based control method for offloading tasks from mobile devices to the cloud,” in Proc. IEEE Global Commun. Conf. (GLOBECOM).

[12] S. Barbarossa et al. “Joint allocation of computation and communication resources in multiuser mobile cloud computing,” in Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC), Darmstadt, Germany, Jun. 2013, pp. 26–30.

[13] Y. Mao et al., “Power-delay tradeoff in multi-user mobile-edge computing systems,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Washington, DC, Dec. 2016, pp. 1–6.

[14] K. Zhang et al., "Energy Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," in IEEE Access, vol. 4, pp. 5896-5907.

[15] Y. Mao, J. Zhang and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 12, pp. 3590-3605, Dec. 2016.

[16] J. Liu et al., "Content caching at the wireless network edge: A distributed algorithm via belief propagation," 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016, pp. 1-6.