Cost Minimization for Cooperative Mobile Edge Computing Systems

Weijian Chen, Yejun He* and Jian Qiao
Guangdong Engineering Research Center of Base Station Antennas and Propagation
Shenzhen Key Laboratory of Antennas and Propagation
College of Electronics and Information Engineering, Shenzhen University, Shenzhen, 518060, China
Email: heyejun@126.com

***Abstract*—Mobile Edge Computing (MEC) is a promising technique that provides distributed computing and storage resources at the edge of a network. In this talk, we investigate a stochastic optimization problem to minimize the cost of MEC system. Firstly, we use the stability of the task buffer queue as constraints to formulate the cost minimization problem. Then we propose the Lyapunov optimization theory to transform the original problem into a deterministic problem. The optimal CPU frequency and optimal transmit power can be obtained in a closed form. In addition, we establish a [O(1/V), O(V)] tradeoff between the system cost and execution latency. Simulation results are provided to verify theoretical analysis and demonstrate the effects of various parameters.**

***Index Terms*—Mobile edge computing, Cost minimization, Lyapunov optimization**

I. Introduction

With the growing explosion of smart devices such as smart phones, tablet computers, and wearable devices, the overall mobile data traffic is expected to increase by more than 46% annually to 77 exabytes per month by 2022 [1]. More and more computation-intensive applications (e.g., augmented reality, virtual reality, face recognition, and interactive online gaming) have emerged and attracted great attention. However, existing mobile devices are generally resource-constrained and cannot process these applications in real time. Therefore, it is necessary to improve the computation performance. Computation offloading is considered as an effective way to improve computation performance of mobile devices.

In conventional cloud computing systems, computation tasks are offloaded to remote cloud servers which would produce huge transmission delay. Different from cloud computing, mobile edge computing (MEC) [2], [3], [4] can provide distributed computing and storage resources at the edge of networks close to mobile users. By offloading computation tasks to MEC servers, the MEC system cannot only reduce the power consumption but also can improve the computation performance.

Some recent studies [5], [6], [7] focus on minimizing system cost or maximizing the profit of MEC servers or making the tradeoff between the system cost and execution delay in the MEC system. In [5], the authors propose an unified optimization framework to maximize the profit of the mobile service provider. In [6], the DJORC algorithm is proposed to maximize the MEC server's economical profit. In [7], based on Lyapunov optimization, the DDROV algorithm is proposed to balance the cost-delay tradeoff in the competition scenario.

In this paper, we consider the problem of system cost minimization in a cooperative scenario. In the cooperative scenario, we consider the situation that a university provides MEC server, and mobile users are students. Then the MEC server will provide the free MEC service to the mobile users. In the MEC system, a part of the computation tasks are computed in the mobile device and the other would be offloaded to MEC servers. In addition, we formulate the stochastic optimization problem which aims to minimize the system cost with the queue stability as constraints. There are several challenges in such a problem. Firstly, the stochastic optimization problem needs the distribution information of the computation task arrival and wireless channel. Secondly, the MEC system needs to minimize the system monetary cost while maintaining the low-latency, which requires the MEC system to make the tradeoff between the two metrics. By leveraging the Lyapunov optimization theory [8], we propose an algorithm to tackle the monetary cost minimization problem.

The rest of this paper is organized as follows. The system models including the computation model, and cost model are given in Section II. Based on the proposed model, an system monetary cost minimization problem is formulated in Section III. Section IV describes the proposed algorithm. Section V discusses the simulation results. At last, Section VI concludes the paper.

II. System Model

As shown in Fig. 1, the MEC system is consisting of a MEC server which is located at the school and the mobile devices. The MEC server provides free services for mobile devices. In the MEC system, the mobile devices are running independent and fine-grained tasks [9]. Time is slotted and the length of the time slot is represented as τ, and the set of the time slot is indexed by $T \triangleq \{0, 1 \cdots\}$. Next, we describe the model of several parts of the edge computing system in Figure.1 in detail.

A. Computing Task and Task Queueing Model

At the beginning of the t-th time slot, mobile device generates a $A(t)$ bits of computation task, with a processing density λ (in CPU cycles/bit). Without loss of generality, we assume

978-1-7281-0660-1/19/$31.00 ©2019 IEEE

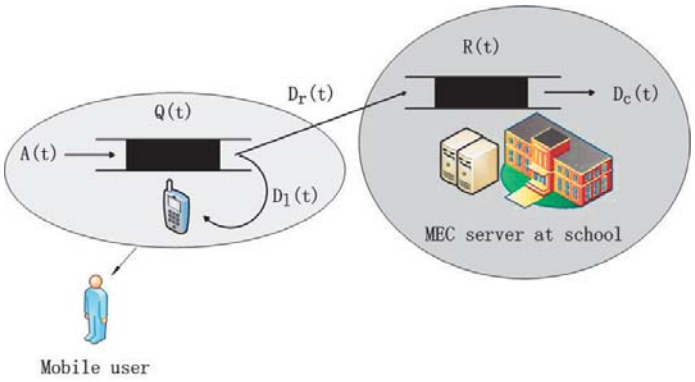

Fig. 1. The MEC system

the input data $A(t)$ is independent and identically distributed (i.i.d.) over time slots and is limited by $0 \leq A(t) \leq A_{\max}$.

In each time slot, a part of the computation tasks will be executed at the mobile device, denoted as $D_l(t)$. The other computation tasks will be offloaded to the MEC server, denoted as $D_r(t)$. The arrived but not executed tasks will be queued in the task buffers at the mobile device. The task buffer of the mobile device is represented as $Q(t)$, and we define $Q(0) = 0$. The evolution of $Q(t)$ can be expressed as

$$Q(t+1) = \max\{Q(t) - D_l(t) - D_r(t), 0\} + A(t). \quad (1)$$

Similarly, the tasks that have been offloaded but not executed by the MEC server will be queued in the task buffers at the MEC server. We denote the task buffer of the MEC server as $R(t)$, and define $R(0) = 0$. Then $R(t)$ evolves according to the following equation

$$R(t+1) = \max\{R(t) - D_c(t), 0\} + D_r(t), \quad (2)$$

where $D_c(t)$ is the amount of computation tasks executed by the MEC server at the t-th time slot.

B. Local computing model and Mobile-edge computing model

In this paper, computation tasks can be processed on mobile devices or offloaded to MEC servers. Next we introduce the local computing model and mobile-edge computing model, respectively.

Local Computing Model: In the local computing model, we denote the CPU-cycle frequency at the t-th time slot as $f_l(t)$, and the maximum allowable CPU-cycle frequency is defined as $f_l^{\max}$, i.e.,

$$0 \leq f_l(t) \leq f_l^{\max}, t \in T. \quad (3)$$

Denote the locally executed computation tasks as $D_l(t)$, it can be expressed as

$$D_l(t) = \tau f_l(t) \lambda^{-1}. \quad (4)$$

Therefore, the power consumption of the mobile device at time slot t denoted as $p_l(t)$ can be given by

$$p_l(t) = \kappa f_l^3(t), \quad (5)$$

where κ is the effective energy coefficient related to the CPU chip architecture [10].

Mobile-edge Computing Model: In the mobile-edge computing model, the input data $A(t)$ should be transferred to the MEC server firstly. Then the computation task will be processed in the MEC server, and the computation results will be transmitted back to the mobile device. Since the size of computation result is generally very small, the transmit process of the computation results is negligible. Let $p(t)$ denote the transmission power of the mobile device, which cannot exceed the maximum value $p_{\max}$, i.e.,

$$0 \leq p(t) \leq p_{\max}, t \in T. \quad (6)$$

According to the Shannon formula, for AWGN channel, the amount of computation task offloaded from the mobile device at tth time slot can be expressed as

$$D_r(t) = \omega \tau log_2(1 + \frac{h(t)p(t)}{\sigma}), \quad (7)$$

where ω is the system bandwidth, σ is the noise power, $h(t)$ is the channel gain.

Thus, the energy consumption in computation offloading can be given by

$$E_r(t) = \tau p(t). \quad (8)$$

After transferring the input data to MEC server, the computation task can be processed by the MEC server. We denote the CPU cycle frequency of the MEC server as $f_c(t)$, which cannot exceed the maximum value $f_c^{\max}$, i.e.,

$$0 \leq f_c(t) \leq f_c^{\max}, t \in T. \quad (9)$$

Then $D_c(t)$ can be expressed as

$$D_c(t) = \tau f_c(t) \lambda^{-1}, \quad (10)$$

Accordingly, we denote the power consumption of the computation task $A(t)$ executed on the MEC server at time slot t as $p_c(t)$, it can be written as

$$p_c(t) = \kappa_{ser} f_c^3(t), \quad (11)$$

where κ_{ser} is the effective energy coefficient at the MEC server related to the CPU chip architecture.

C. Cost Model

In this section, we will discuss the monetary cost of the MEC system in the cooperative scenario, and define the system cost as the sum of the monetary cost of mobile device and MEC server.

On the user side, the mobile user needs to pay the following types of monetary costs: energy cost for computing task locally and energy cost for transferring the input data to the MEC server. On the MEC server side, the MEC servers need to pay the following types of monetary costs: electricity cost for computing task, cost for renting radio bandwidth from network operators. In this paper, we consider the system cost including the cost of user side and the cost of MEC server side ,and denoted as $U(t)$, it can be given by

$$U(t) = \alpha E_l(t) + \alpha E_r(t) + \alpha E_c(t) + \delta D_r(t), \quad (12)$$

where $E_l(t), E_c(t)$ is the energy consumption of the mobile device and the MEC server, respectively. α (in \$/J) is a weight parameter which transforms energy consumption into money and depends on the human sensitiveness on money and energy consumption [11], δ (in \$/bit) is the price for renting radio bandwidth from network operators.

III. Problem Formulation

In this paper, we construct the optimization problem where the objective is to minimize the time-averaged system cost with the queue stability constraints of both user side and MEC server side. Thus the optimization problem can be formulated as

$$\begin{aligned}\mathcal{P}_1: &\min_{f_l(t),p(t),f_c(t)} \lim_{T\to+\infty} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{U(t)\}\\ &s.t. (3),(6),(9)\\ &\lim_{T\to+\infty}\frac{1}{T}\mathbb{E}\{Q(t)+R(t)\}<\infty \end{aligned} \tag{13}$$

where (3) is the CPU-cycle frequency constraint of the mobile device, (6) is the transmit power constraint of the mobile device, (9) is the CPU-cycle frequency constraint of the MEC server, (13) is the queue stability constraint.

IV. Online Computation Offloading Algorithm

In this section, we propose an online computation offloading algorithm based on Lyapunov optimization theory to tackle the problem of minimizing system cost, and then analyze the performance of the proposed algorithm.

A. The OCO Algorithm

First, we define the quadratic Lyapunov function as

$$L[\Theta(t)] \triangleq \frac{1}{2}[Q^2(t)+R^2(t)], \tag{14}$$

where $\Theta(t) \triangleq [Q(t), R(t)]$. Then the Lyapunov drift function $\Delta[\Theta(t)]$ can be represented as

$$\Delta[\Theta(t)] = \mathbb{E}\{L[\Theta(t+1)] - L[\Theta(t)] \mid \Theta(t)\} \tag{15}$$

Based on Lyapunov optimization theory, in order to minimize the objective function in $\mathcal{P}_1$, the Lyapunov drift-plus-penalty function is defined as

$$\Delta_V[\Theta(t)] = \Delta[\Theta(t)] + V\mathbb{E}[U(t) \mid \Theta(t)], \tag{16}$$

where V (in $bit^2/\$$) is a non-negative control parameter that determines a tradeoff between system cost and system queue length.

For ease of analysis, we derive the upper bound for $\Delta_V(\Theta(t))$, which is given by

$$\begin{aligned}\Delta_V(\Theta(t)) \le C &+ \mathbb{E}\{Q(t)[A(t)-D_l(t)-D_r(t)]\}\\ &+ \mathbb{E}\{R(t)[D_r(t)-D_c(t)\} + V\mathbb{E}[U(t)],\end{aligned} \tag{17}$$

where $C = \frac{1}{2}[A_{\max}^2 + (D_l^{\max})^2 + 2(D_r^{\max})^2 + (D_c^{\max})^2]$.

Proof: The proof is omitted due to space limitation. ■

Therefore, we convert the original optimization problem into the problem of minimize the upper bound of $\Delta_V(\Theta(t))$ in the right-hand side of (17) at each time slot. And based on the expression of $\Delta_V(\Theta(t))$, we proposed an online computation offloading algorithm to tackle this deterministic optimization problem at each time slot, which is summarized in *Algorithm 1*. Note that the objective function of $\mathcal{P}_2$ corresponds to the right-hand side of (17), where $A(t)$ can be viewed as a constant and obtained at the beginning of each time slot, and all the constraints in $\mathcal{P}_1$ except the task buffer stability constraint in (13) are retained in $\mathcal{P}_2$. The optimal solution for $\mathcal{P}_2$ will be developed in the next subsection.

Algorithm 1 Online Computation Offloading Algorithm

1: Set $t=0, Q(0)=0, R(0)=0$;
2: **while** $t<T$ **do**
3: At each time slot, obtain $A(t), Q(t), R(t)$.
4: Determine $f_l(t), p(t), f_c(t)$ by solving

$$\begin{aligned}\mathcal{P}_2: \min_{f_l(t),p(t),f_c(t)} &-Q(t)[D_l(t)+D_r(t)] + VU(t)\\ &+R(t)[D_r(t)-D_c(t)]\\ s.t. \quad &(3),(6),(9)\end{aligned}$$

5: Update $Q(t)$ and $R(t)$ according to (1) and (2).
6: $t \leftarrow t+1$.
7: **end while**

B. Optimal Solution for P2

In this section, we will obtain the optimal CPU frequency and optimal transmit power of the mobile device, and the optimal CPU frequency of the MEC server by solving $\mathcal{P}_2$.

Optimal CPU frequency of the mobile device: After decoupling $f_l(t)$ from $\mathcal{P}_2$, the optimal value of the $f_l(t)$ can be obtained by solving the following problem

$$\min_{0\le f_l(t)\le f_l^{\max}} -Q(t)\tau f_l(t)\lambda^{-1} + \alpha\kappa\tau V f_l^3(t), \tag{18}$$

and its closed form optimal solution can be expressed as

$$f_l^*(t) = \min\{f_l^{\max}, \sqrt{\frac{Q(t)}{3\alpha\kappa\lambda V}}\}. \tag{19}$$

Optimal transmit power: After decoupling $p(t)$ from $\mathcal{P}_2$, the optimal value of $p(t)$can be obtained by solving

$$\min_{0\le p(t)\le p_{\max}} [V\delta + R(t) - Q(t)]\tau\omega\log_2(1+\frac{p(t)h(t)}{\sigma^2}) + \alpha p(t)\tau V. \tag{20}$$

Deriving the objective function to $p(t)$ and rearranging the term, then the optimal solution of $p(t)$ can be expressed as

$$p^*(t) = \min\{\max\{\frac{[Q(t)-R(t)-V\delta]\omega}{\alpha V ln2} - \frac{\sigma^2}{h(t)}, 0\}, p_{\max}\}. \tag{21}$$

Optimal CPU frequency of the MEC server: Similarly, after decoupling $f_c(t)$ from $\mathcal{P}_2$, the optimal value of $f_c(t)$ can be obtained by solving:

$$\min_{0\le f_c(t)\le f_c^{\max}} -R(t)\tau f_c(t)\lambda^{-1} + \alpha\tau\kappa_{ser} V f_c^3(t), \tag{22}$$

and its optimal solution can be expressed as

$$f_c^*(t) = \min\{f_c^{\max}, \sqrt{\frac{R(t)}{3\alpha\kappa\lambda V}}\}. \quad (23)$$

C. Performance Analysis

In this section, we present the main numerical results for theoretical analysis of this paper based on Lyapunov optimization theory, which shows the time-averaged system cost and the time-averaged queue backlog of the task buffers can be upper-bounded by the Theorem 1 when the proposed algorithm is adopted.

Theorem 1: For the system defined in section II, when adopting the proposed algorithm, the time-averaged system cost satisfies

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{U(t)\} \le U^* + \frac{C}{V}, \quad (24)$$

where U^* is the optimal value of $\mathcal{P}_1$. Supposed there are $\epsilon > 0$ and $U(\epsilon)$ which satisfy the Slater conditions [8], then the sum of time-averaged queue length of the task buffers of mobile device and MEC server satisfies

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{Q(t)+R(t)\} \le \frac{C + V(U(\epsilon) - U^*)}{\epsilon}. \quad (25)$$

Proof: The proof is omitted due to space limitation. ■

Remark 1: According to Little's Law [12], the execution delay is proportional to the time-averaged queue length of the task buffers. Thus Theorem 1 shows that there exists an $[O(1/V), O(V)]$ tradeoff between the time-averaged system cost and the sum of time-averaged queue length of task buffers of the mobile device and MEC server. When the proposed online algorithm is adopted, the time-averaged system cost decreases inversely proportional to V, while the sum of time-averaged queue length of task buffers of mobile device and MEC server increases with V. Therefore, we can adjust V to balance these two metrics.

V. SIMULATION RESULTS

In this section, we use simulations to validate the theoretical analysis and evaluate the effects of the system parameter. In the simulation, we assume the distance between mobile device and MEC server is 50 meters away. The channel power gains $h(t)$ are exponentially distributed with mean $g_0(d_0/d)^4$, where $g_0 = -40$ dB and $d_0 = 1$ m [13]. In addition, we set $\kappa = \kappa_{ser} = 10^{-27}, \alpha = 2.44 \times 10^{-4}$ \$/J [11], $\delta = 0.5 \times 10^{-10}$ \$/bit, $\tau = 1$ ms, $\omega = 1$ MHz, $\sigma = 10^{-13}$ W, $p_{\max} = 1$ W, $f_{\max} = 1.5$ GHz, $\lambda = 737.5$ cycles per byte, $A_{\max} = 1000$ bits. The simulation results are averaged over 10000 time slots.

First, we show the tradeoff between the time-averaged system cost and the sum of time-averaged queue length of task buffers of the mobile device and MEC server in Fig. 2. As can be seen, the blue curve in Fig. 2 depicts the relationship between the average system cost and the parameter V, it shows that the system cost decrease as V increase and converges to U^* when V is sufficiently large. And the sum of time-averaged

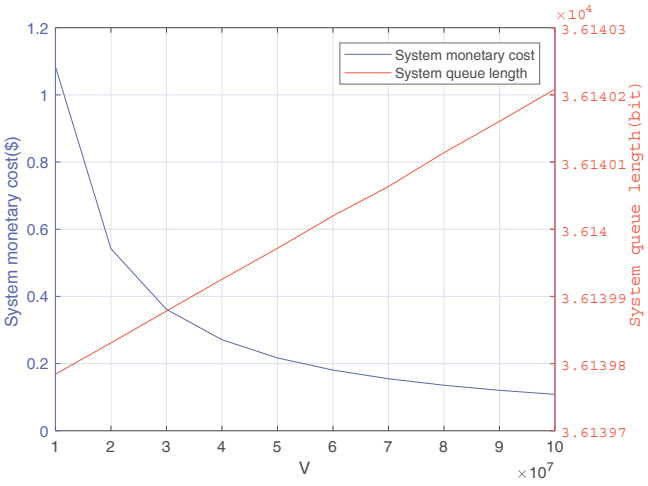

Fig. 2. Average system cost queue length vs. the parameter V

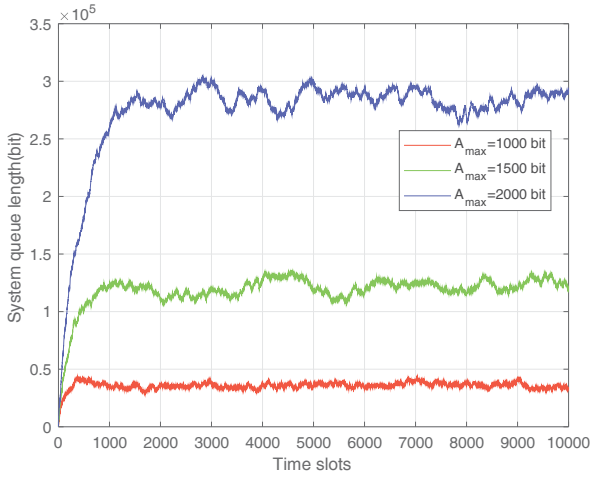

Fig. 3. Length of Q+R vs. time ($V = 2 \times 10^6\ bit^2/\$$)

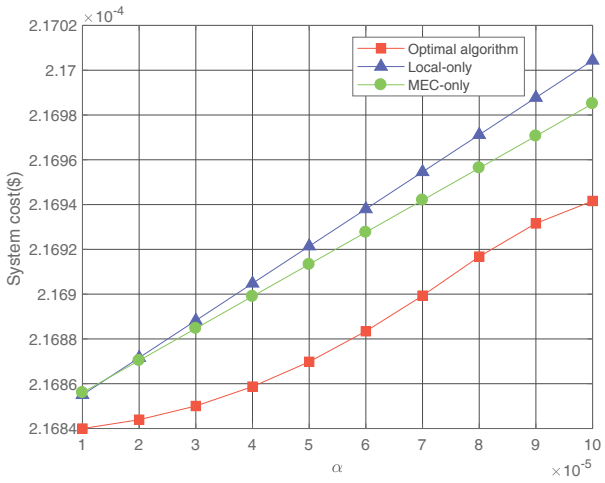

Fig. 4. System cost vs. the parameter α

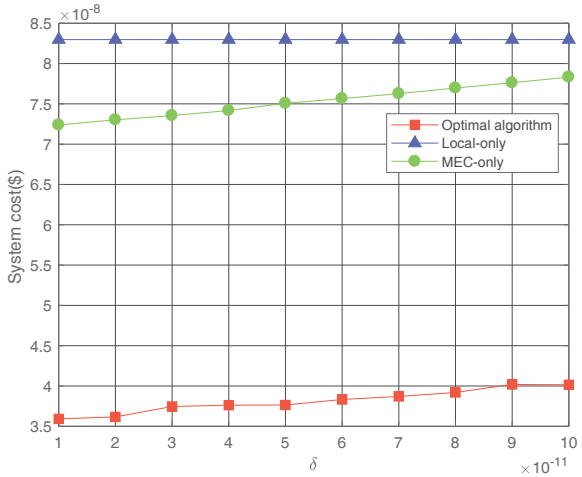

Fig. 5. System cost vs. the parameter δ

queue length of task buffers of the mobile device and MEC server are increasing as V increase as shown in Fig. 2. Thus, it can be known that the balance between the time-averaged system cost and the sum of time-averaged queue length of task buffers of the mobile device and MEC server can be achieved by adjusting the parameter V.

Next, we show the stability of the queue length of task buffers of the MEC system, as shown in Fig. 3. As can be seen, for all three cases, the queue length of the system increases at the beginning, and stabilizes around 3.7×10^4 bits, 1.2×10^5 bits, 2.8×10^5 bits respectively. Thus it demonstrates the stability of the MEC system. In addition, it also shows that the queue length of the system increases as the input data increase.

Last, we compare our proposed algorithm with Local-only and MEC-only algorithms to show the performance of our proposed algorithm. In Local-only, the computation tasks only executed in the mobile device by the maximum CPU frequency, and in MEC-only, the input data of the computation tasks will be transferred to the MEC server with the maximum transmit power and then executed in the MEC server by the maximum CPU frequency. As shown in Fig. 4, the system cost increases with the price that is energy-monetary transform parameter, i.e., α. Similarly, as can be seen in Fig. 5, the system cost increases with the price that means the MEC system rents radio bandwidth from network operators, i.e., δ. It is noted that the Local-only algorithm is not affected by the parameter δ since it does not need to use the radio bandwidth.

VI. CONCLUSIONS

In this paper, we investigated the problem of the monetary cost minimization in mobile-edge computing (MEC) system in a cooperative scenario. Then we proposed an online dynamic computation offloading algorithm that correctly decides the local execution and computation offloading policy. Simulation results verify the theoretical analysis, and show that the proposed algorithm can balance the system cost and the queue length of the system.

ACKNOWLEDGEMT

This work was supported in part by the National Natural Science Foundation of China under Grants 61372077, 61801299, and 61871433, in part by the Shenzhen Science and Technology Programs under Grants ZDSYS 201507031550105, JCYJ 20170302150411789, JCYJ 20170302142515949, GCZX 2017040715180580 and GJHZ 20180418190529516, and in part by the Guangzhou Science and Technology Program under Grant 201707010490.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022, 2018.
[2] K. Zhang, S. Leng, Y. He, S. Maharjan and Y. Zhang,"Cooperative Content Caching in 5G Networks with Mobile Edge Computing," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 80-87, June 2018.
[3] K. Zhang, S. Leng, Y. He, S. Maharjan and Y. Zhang, "Mobile Edge Computing and Networking for Green and Low-Latency Internet of Things, " *IEEE Communications Magazine*, vol. 56, no. 5, pp. 39-45, May 2018.
[4] X. Huang, R. Yu, J. Kang, Y. He and Y. Zhang, "Exploring Mobile Edge Computing for 5G-Enabled Software Defined Vehicular Networks, " *IEEE Wireless Communications*, vol. 24, no. 6, pp. 55-63, Dec. 2017.
[5] X. Wang et al., "Dynamic Resource Scheduling in Mobile Edge Cloud with Cloud Radio Access Network, " *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 11, pp. 2429-2445, 1 Nov. 2018.
[6] J. Du, L. Zhao, J. Feng, X. Chu and F. R. Yu, "Economical Revenue Maximization in Cache Enhanced Mobile Edge Computing, " in Proc. IEEE International Conference on Communications (ICC), Kansas City, MO, 2018, pp. 1-6.
[7] J. Du, F. R. Yu, X. Chu, J. Feng and G. Lu, "Computation Offloading and Resource Allocation in Vehicular Networks Based on Dual-Side Cost Minimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1079-1092, Feb. 2019.
[8] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems," *Synthesis Lectures on Communication Networks*, vol. 1, no. 1, pp. 1-211, 2010.
[9] Y. Mao, J. Zhang, S. H. Song , et al., "Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems," *IEEE Trans. Wireless Commun*, vol. 16, no. 9, pp. 5994-6009, 2017.
[10] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *J. VLSI Signal Process. Syst*, vol. 13, no. 2-3, pp. 203-221, Aug. 1996.
[11] Y. Kim, J. Kwak, and S. Chong, "Dual-side optimization for cost-delay tradeoff in mobile edge computing," *IEEE Trans. Veh. Technol*, vol. 67, no. 2, pp. 1765-1781, 2018.
[12] S. M. Ross, Introduction to probability models. Academic press, 2014.
[13] 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.