Joint User Association and Resource Allocation Optimization for MEC-Enabled IoT Networks

Yaping Sun*, Jie Xu*, and Shuguang Cui*†, *Fellow, IEEE*
*SSE and FNii, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
†Shenzhen Research Institute of Big Data, Shenzhen, China
Email: sunyaping@cuhk.edu.cn, xujie@cuhk.edu.cn, shuguangcui@cuhk.edu.cn

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9839276

***Abstract*—This paper studies a mobile edge computing (MEC) network to support emerging Internet-of-things applications, where multiple access points (APs), each attached with an MEC server, need to collect data from multiple sensors, process them, and then send computation results to the paired actuators for control. Specifically, we consider a three-phase operation protocol for data uploading, edge computing, and results downloading, where the frequency-division multiple access is implemented to accommodate communications of multiple sensors/actuators. Under this setup, we minimize the end-to-end (E2E) latency of the sensing-communication-computation-actuation loop by properly designing the user association and resource allocation policy, subject to the communication and computation resource constraints. The formulated problem, however, is a mixed-integer non-linear program that is difficult to be optimized. Despite this fact, we first search the optimal solution for user association, and then apply convex optimization for resource allocation given the obtained optimal user association. Finally, numerical results show that the proposed optimal joint design significantly reduces the E2E latency, as compared to conventional designs without such joint designs.**

***Index Terms*—Mobile edge computing (MEC), sensing-communication-computation-actuation loop, joint communication and computation resource allocation, user association.**

I. Introduction

Recent technical advancements in Internet of things (IoT) and artificial intelligence (AI) have enabled various emerging applications (e.g., industrial automation, auto-driving, and virtual reality (VR)) with extensive communication and computation requirements. How to ensure an ultra-low-latency sensing-communication-computation-actuation service loop becomes an essential challenge faced in the design of wireless networks [1]. For instance, in order to support auto-driving, sensors (e.g., cameras on the cars or on the roadside) need to sense the environment to obtain traffic information, which is then processed in a swift way to facilitate the vehicles' auto-driving. In general, these applications demand ultra-high communication rates (e.g., on the order of several Gigabits per second) and ultra-low end-to-end (E2E) service latency (e.g., on the order of several milliseconds) for sensing, communication, computation, and actuation, which are becoming great burdens for mobile operators [2].

To cope with such challenges, mobile edge computing (MEC) has emerged as one promising solution via pushing the computation resources to the access points (APs)/base stations (BSs) at the wireless network edge close to end users [3]. With MEC, source devices (e.g., sensor nodes) can first offload the data or task input bits to the MEC servers at APs/BSs for remote computation, and then the APs/BSs can send the computation results to destination devices (e.g., actuator nodes) for command and control. As a result, the MEC technique can significantly reduce the E2E service latency, and decrease the traffic loads in core networks [3].

For instance, some early works [4], [5] focused on the uplink data/computation offloading in MEC networks, in which the results downloading time is assumed to be negligible or constant, by assuming the computation results to be with relatively small sizes and/or the transmit power at the APs to be relatively large [4], [5]. Notice that as these works assumed negligible computation results and ignored the downloading process design for simplicity, they are not applicable for applications that are bandwidth hungry in downloading the computation results, such as image processing and mobile VR video delivery [6]. In the other line of research, some prior works [7]–[11] investigated the MEC networks with both uplink and downlink transmission, by considering single-server [7] and multi-server MEC systems [8]–[11], respectively. When there are multiple MEC servers, the user association problem between multiple devices and multiple MEC servers has been investigated in [8]–[11]. Notice that as [8] and [9] only considered single-user scenarios, they are not applicable for multi-user MEC systems with limited communication and computation resources. When there are multiple users, [10] and [11] did not consider the fairness-aware resource allocation among the multiple users, which is essential for further enhancing the MEC performance. This thus motivates the investigation in this work.

This paper considers a low-latency MEC network to support emerging IoT applications, which consists of multiple APs/MEC servers, and a set of distributed sensor-actuator pairs. In this network, each sensor needs to offload its sensing data to one of the APs for remote processing, and then the AP sends the computation results to the correspondingly paired actuators. Due to the limited communication and computation

The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Shenzhen Outstanding Talents Training Fund 202002, Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, the National Natural Science Foundation of China under grants Nos. U2001208 and 61871137, the Science and Technology Program of Guangdong Province under grant No. 2021A0505030002, and the Shenzhen Fundamental Research Program under grant No. 20210318123512002.

resources distributed in this network, how to properly associate the sensor-actuator pairs with APs and efficiently allocate the communication and computation resources among them for minimizing the E2E latency of the sensing-communication-computation-actuation loop is an important but challenging task. On one hand, different from prior works considering the source and destination nodes are co-located, this paper considers practical IoT networks in which the paired sensors and actuators may be placed at different locations but need to be associated with the same AP to minimize the service latency. In this case, the sensors and actuators may have distinct wireless channel qualities with each AP, thus making the user association designs in [8]–[11] not applicable. On the other hand, the user association design in such MEC-enabled IoT networks needs to further consider the computation load balance among different APs, in addition to their communication loads. Due to the heterogeneous computation and communication capabilities at APs, the user association and resource allocation problem under our setup will become more difficult to handle.

To tackle the aforementioned issues, we consider a three-phase operation protocol for data uploading, edge computing, and results downloading, respectively, and implement the frequency-division multiple access (FDMA) to accommodate the communications of multiple sensors/actuators. Our objective is to minimize the E2E latency of the sensing-communication-computation-actuation service loop, which is defined as the sum of the uplink transmission latency, the computation latency, and the downlink transmission latency. Towards this end, we jointly optimize the user association (between sensor/actuator pairs and APs), transmit bandwidth and power allocation, computation frequency allocation, and the time durations for the three phases, subject to the distributed communication and computation resource constraints. The formulated minimization problem is a mixed-integer non-convex program and thus very difficult to be optimized. To solve it, we use the brute-force search to find the optimal sensor/actuator-AP association, under each of which convex optimization techniques are employed to optimally allocate the communication and computation resources. Numerical results show the promising performance gains of the optimal algorithm, as compared with conventional designs without such joint optimization under different system parameters.

II. System Model and Problem Formulation

A. System Model

This paper considers an MEC-enabled IoT network as shown in Fig. 1, which consists of K sensors, M APs each integrated with an MEC server, and L actuators. We denote the sets of sensors, APs and actuators as $\mathcal{K} \triangleq \{1, \cdots, K\}$, $\mathcal{M} \triangleq \{1, \cdots, M\}$, and $\mathcal{L} \triangleq \{1, \cdots, L\}$, respectively. In the MEC-enabled IoT network, the APs need to collect the sensing information from each sensor, process the messages, and then send the computation results (or the command and control signals) to the correspondingly paired actuator. Suppose that each actuator $l \in \mathcal{L}$ is paired with sensor $\alpha(l) \in \mathcal{K}$, i.e., the computation results based on the sensing data from sensor $\alpha(l)$ are requested by actuator l. Notice that the pairing between actuator l and sensor $\alpha(l)$ is determined *a priori* based on specified applications. Taking auto-driving as an example [12], the actuator-sensor pairs, i.e., vehicle-camera/radar pairs, can be determined based on the vehicles' navigation. The computation task for each sensor $k \in \mathcal{K}$ is characterized by a three-tuple $\{I_k,\ W_k,\ O_k\}$, where I_k (in bits) denotes the size of the sensing data generated by sensor k, W_k (in cycles/bit) denotes the CPU cycles required for computing one bit of this task, and O_k (in bits) denotes the size of the output data.

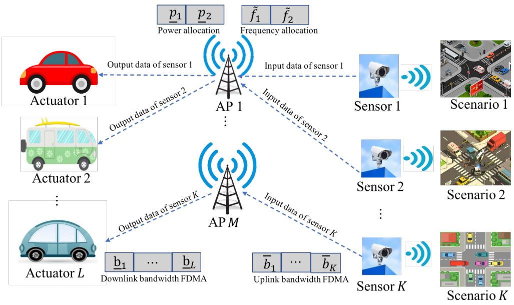


Fig. 1: Illustration of the MEC-enabled IoT network.

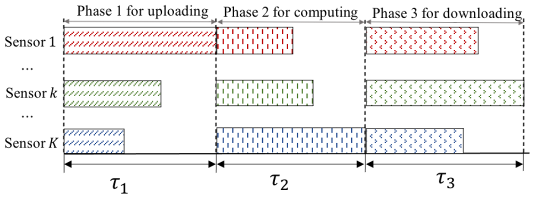


Fig. 2: Service protocol for the MEC-enabled IoT network, in which $\mathcal{K} = \mathcal{L}$ and $\alpha(l) = l, \forall l \in \mathcal{L}$.

In order to execute the computation tasks in a swift manner, we consider that each sensor-actuator pair needs to associate with one AP to avoid the data transmission over the backhaul network among different APs. Let $x_{k,m} \in \{0, 1\}$ denote the association indicator between AP $m \in \mathcal{M}$ and sensor $k \in \mathcal{K}$ (and accordingly actuator l with $\alpha(l) = k$), where $x_{k,m} = 1$ indicates that sensor k is associated with AP m, and $x_{k,m} = 0$ otherwise. By considering that each sensor is associated with only one AP, we have $\sum_{m=1}^{M} x_{k,m} = 1, \forall k \in \mathcal{K}$.

We consider the service protocol for this MEC network as shown in Fig. 2, which consists of three phases, i.e., Phase 1 for task offloading from sensors to APs, Phase 2 for remote task execution at APs, and Phase 3 for command/control signal broadcasting from APs to actuators. Let τ_1, τ_2, and τ_3 denote the time duration for uploading, computing, and broadcasting, respectively. Furthermore, we consider the FDMA protocol for both task uploading from multiple sensors to APs and downloading from APs to multiple actuators to avoid the potential co-channel interference. Notice that the design principles in this paper are generally extendable to scenarios allowing frequency reuse among different APs by properly controlling their interference (e.g., via power control and interference mitigation), which can be left for future work. In the following, we consider the three phases, respectively.

First, we consider the FDMA-based task offloading from the sensors to the APs in Phase 1. Let $\bar{b}_k \in [0, B]$ denote the bandwidth allocated to sensor k for uploading. Thus, we have $\sum_{k=1}^{K} \bar{b}_k \leq B$. The achievable uplink data rate from sensor k to AP m is given by

$$\bar{R}_{k,m}(\bar{b}_k) = \bar{b}_k \log_2 \left(1 + \frac{\bar{P}_k \bar{g}_{k,m}}{\bar{b}_k N_0}\right), \ \forall k \in \mathcal{K}, m \in \mathcal{M} \quad (1)$$

where $\bar{P}_k$ denotes the fixed transmit power of sensor k, $\bar{g}_{k,m}$ denotes the channel power gain from sensor k to AP m, and N_0 denotes the noise power spectral density. In order to ensure the input data from sensor k to be successfully uploaded to the associated MEC server within duration τ_1, we have $\frac{I_k}{\tau_1} \leq \sum_{m=1}^{M} x_{k,m} \bar{b}_k \log_2 \left(1 + \frac{\bar{P}_k \bar{g}_{k,m}}{\bar{b}_k N_0}\right), \forall k \in \mathcal{K}$.

Next, after uploading in Phase 1, the APs execute the tasks in Phase 2 via network function virtualization (NFV), where each MEC server allocates different virtual machines for executing different tasks simultaneously [3]. Let F_m (in cycles/s) denote the maximum available computation frequency of the MEC server at AP m, and $\tilde{f}_k$ (in cycles/s) denote the computation frequency allocated for sensor k. Thus, we have $\sum_{k=1}^{K} x_{k,m} \tilde{f}_k \leq F_m, \forall m \in \mathcal{M}$ [8]. In order to guarantee that the task computation for each sensor $k \in \mathcal{K}$ is finished within duration τ_2, we have $\frac{I_k W_k}{\tilde{f}_k} \leq \tau_2, \forall k \in \mathcal{K}$.

Then, we consider Phase 3 in which the APs transmit the output data to the corresponding actuators based on the FDMA. Let $\underline{b}_l \in [0, B]$ denote the bandwidth allocated to actuator l for receiving the command/control signals from its associated AP. Thus, we have $\sum_{l=1}^{L} \underline{b}_l \leq B$. Notice that each AP needs to serve multiple actuators by allocating the transmit power among them. Let $\underline{p}_l \geq 0$ denote the transmit power for actuator l. Suppose that the maximum transmit power of AP m is $\underline{P}_m$, $\forall m \in \mathcal{M}$. Then, we have $\sum_{l=1}^{L} x_{\alpha(l),m} \underline{p}_l \leq \underline{P}_m$, $\forall m \in \mathcal{M}$. Let $\underline{g}_{m,l}$ denote the channel power gain from AP m to actuator l. The achievable downlink data rate from AP m to actuator l is given by

$$\underline{R}_{m,l}\left(\underline{b}_l, \underline{p}_l\right) = \underline{b}_l \log_2 \left(1 + \frac{\underline{p}_l \underline{g}_{m,l}}{\underline{b}_l N_0}\right), \ \forall m \in \mathcal{M}, l \in \mathcal{L}. \quad (2)$$

In order to ensure the output data with size $O_{\alpha(l)}$ to be successfully downloaded from the AP to actuator l within duration τ_3, we have $\frac{O_{\alpha(l)}}{\tau_3} \leq \sum_{m=1}^{M} x_{\alpha(l),m} \underline{b}_l \log_2 \left(1 + \frac{\underline{p}_l \underline{g}_{m,l}}{\underline{b}_l N_0}\right), \forall l \in \mathcal{L}$.

By combining the above three phases, the E2E service latency for completing all the computation tasks is given by $T_{E2E} = \tau_1 + \tau_2 + \tau_3$.

B. Problem Formulation

Under this setup, our objective is to minimize the E2E service latency T_{E2E} of these sensor-actuator pairs, by jointly optimizing their association with APs $\{x_{k,m}\}$, uplink transmit bandwidth allocation $\{\bar{b}_k\}$, computation allocation $\{\tilde{f}_k\}$, downlink transmit bandwidth $\{\underline{b}_l\}$ and power allocation $\{\underline{p}_l\}$, as well as time duration allocation $\{\tau_1, \tau_2, \tau_3\}$. Let $\boldsymbol{x}$ denote a $KM \times 1$ vector collecting $\{x_{k,m}\}$, $\bar{\boldsymbol{b}}$ denote a $K \times 1$ vector collecting $\{\bar{b}_k\}$, $\tilde{\boldsymbol{f}}$ denote a $K \times 1$ vector collecting $\{\tilde{f}_k\}$, $\underline{\boldsymbol{b}}$ denote a $L \times 1$ vector collecting $\{\underline{b}_l\}$, $\underline{\boldsymbol{p}}$ denote a $L \times 1$ vector collecting $\{\underline{p}_l\}$, and $\boldsymbol{\tau}$ denote a 3×1 vector collecting $\{\tau_1, \tau_2, \tau_3\}$. The optimization problem is formulated as

$$\begin{aligned}
\text{(P1)}: \min_{\boldsymbol{x}, \bar{\boldsymbol{b}}, \tilde{\boldsymbol{f}}, \underline{\boldsymbol{b}}, \underline{\boldsymbol{p}}, \boldsymbol{\tau}} \quad & \tau_1 + \tau_2 + \tau_3 \\
\text{s.t.} \quad & \sum_{m=1}^{M} x_{k,m} = 1, \forall k \in \mathcal{K} && (3) \\
& \sum_{k=1}^{K} \bar{b}_k \leq B && (4) \\
& \frac{I_k}{\tau_1} \leq \sum_{m=1}^{M} x_{k,m} \bar{b}_k \log_2 \left(1 + \frac{\bar{P}_k \bar{g}_{k,m}}{\bar{b}_k N_0}\right), \forall k \in \mathcal{K} && (5) \\
& \sum_{k=1}^{K} x_{k,m} \tilde{f}_k \leq F_m, \forall m \in \mathcal{M} && (6) \\
& \frac{I_k W_k}{\tilde{f}_k} \leq \tau_2, \forall k \in \mathcal{K} && (7) \\
& \sum_{l=1}^{L} \underline{b}_l \leq B && (8) \\
& \sum_{l=1}^{L} x_{\alpha(l),m} \underline{p}_l \leq \underline{P}_m, \forall m \in \mathcal{M} && (9) \\
& \frac{O_{\alpha(l)}}{\tau_3} \leq \sum_{m=1}^{M} x_{\alpha(l),m} \underline{b}_l \log_2 \left(1 + \frac{\underline{p}_l \underline{g}_{m,l}}{\underline{b}_l N_0}\right), \forall l \in \mathcal{L} && (10) \\
& x_{k,m} \in \{0, 1\}, \forall k \in \mathcal{K}, m \in \mathcal{M} && (11) \\
& \bar{b}_k \geq 0, \tilde{f}_k \geq 0, \forall k \in \mathcal{K} \\
& \underline{b}_l \geq 0, \underline{p}_l \geq 0, \forall l \in \mathcal{L} \\
& \tau_i \geq 0, \ \forall i \in \{1, 2, 3\}.
\end{aligned}$$

Problem (P1) is a combinatorial mixed-integer non-linear problem due to the binary constraints in (11), which is thus very difficult to be optimally solved.

III. Optimal Solution to Problem (P1)

In this section, we present the optimal solution to (P1) based on the brute-force search. We first solve for $\bar{\boldsymbol{b}}$, $\tilde{\boldsymbol{f}}$, $\underline{\boldsymbol{b}}$, $\underline{\boldsymbol{p}}$, and $\boldsymbol{\tau}$ under any given $\boldsymbol{x}$, and then search over $\boldsymbol{x}$ via the brute-force search. First, under any given $\boldsymbol{x}$, (P1) can be equivalently decomposed into the following three parallel subproblems (P2), (P3), and (P4) for uplink bandwidth allocation, computation allocation, as well as downlink bandwidth and power allocation optimization, respectively.

$$\begin{aligned}
\text{(P2)}: \min_{\{\bar{b}_k \geq 0\}, \tau_1 \geq 0} \quad & \tau_1 \\
\text{s.t.} \quad & \sum_{k=1}^{K} \bar{b}_k \leq B && (12) \\
& \frac{I_k}{\tau_1} \leq \bar{b}_k \log_2 \left(1 + \frac{\bar{P}_k \bar{g}_{k,m_k(\boldsymbol{x})}}{\bar{b}_k N_0}\right), \forall k \in \mathcal{K} && (13)
\end{aligned}$$

where $m_k(\boldsymbol{x})$ denotes the AP associated with sensor k under given $\boldsymbol{x}$, i.e., $x_{k,m_k(\boldsymbol{x})}=1, \forall k \in \mathcal{K}$.

$$\begin{aligned} \text{(P3)}: \min_{\{\tilde{f}_k \geq 0\}, \tau_2 \geq 0} \quad & \tau_2 \\ \text{s.t.} \quad & \sum_{k=1}^{K} x_{k,m} \tilde{f}_k \leq F_m, \forall m \in \mathcal{M} && (14) \\ & \frac{I_k W_k}{\tilde{f}_k} \leq \tau_2, \forall k \in \mathcal{K}. && (15) \end{aligned}$$

$$\begin{aligned} \text{(P4)}: \min_{\{0 \leq \underline{b}_l \leq B\}, \{\underline{p}_l \geq 0\}, \tau_3 \geq 0} \quad & \tau_3 \\ \text{s.t.} \quad & \sum_{l=1}^{L} \underline{b}_l \leq B && (16) \\ & \sum_{l=1}^{L} x_{\alpha(l),m} \underline{p}_l \leq \underline{P}_m, \forall m \in \mathcal{M} && (17) \\ \frac{O_{\alpha(l)}}{\tau_3} \leq \underline{b}_l \log_2 & \left(1+\frac{\underline{p}_l \underline{g}_{m_{\alpha(l)}(\boldsymbol{x})}}{\underline{b}_l N_0}\right), \forall l \in \mathcal{L}. && (18) \end{aligned}$$

As compared to the original problem (P1), in problem (P4) the additional constraints of $\underline{b}_l \leq B, \forall l \in \mathcal{L}$ are added to facilitate the solution, without loss of optimality.

Suppose that the optimal solutions to the joint communication and computation allocation problems (P2), (P3) and (P4) are denoted by $\left(\bar{\boldsymbol{b}}^*(\boldsymbol{x}), \tilde{\boldsymbol{f}}^*(\boldsymbol{x}), \underline{\boldsymbol{b}}^*(\boldsymbol{x}), \underline{\boldsymbol{p}}^*(\boldsymbol{x}), \boldsymbol{\tau}^*(\boldsymbol{x})\right)$, and the correspondingly obtained minimum E2E latency is given by $T^*_{E2E}(\boldsymbol{x}) = \tau_1^*(\boldsymbol{x}) + \tau_2^*(\boldsymbol{x}) + \tau_3^*(\boldsymbol{x})$. Then, the optimal association policy $\boldsymbol{x}^*$ is obtained via the brute-force search, i.e., $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} T^*_{E2E}(\boldsymbol{x})$, s.t. (3) and (11). In the following, we focus on solving (P2), (P3), and (P4).

A. Optimal Uplink Bandwidth Allocation Solution to (P2)

First, we consider problem (P2). Notice that $\bar{R}_{k,m}(\bar{b}_k) = \bar{b}_k \log_2\left(1+\frac{\bar{P}_k \bar{g}_{k,m}}{\bar{b}_k N_0}\right)$ is a concave function with respect to (w.r.t.) $\bar{b}_k > 0$, and as a result, constraint (13) is convex [13]. Therefore, problem (P2) is convex. It is thus easy to show that at the optimal solution to problem (P2), the following conditions must hold:

$$\sum_{k=1}^{k} \bar{b}_k = B \tag{19}$$

$$\frac{I_k}{\tau_1} = \bar{b}_k \log_2\left(1+\frac{\bar{P}_k \bar{g}_{k,m_k(\boldsymbol{x})}}{\bar{b}_k N_0}\right), \forall k \in \mathcal{K}. \tag{20}$$

By combining (19) and (20), we have the following proposition, for which the proof is omitted for brevity.

Proposition 1. The optimal solution to (P2) is given by

$$\bar{b}_k^*(\boldsymbol{x}) = \frac{\ln 2 I_k}{\tau_1^*(\boldsymbol{x})\left(-\mathcal{W}(-a_k e^{-a_k}) - a_k\right)}, \; \forall k \in \mathcal{K} \tag{21}$$

where $a_k \triangleq \frac{\ln 2 I_k N_0}{\tau_1^*(\boldsymbol{x}) \bar{P}_k \bar{g}_{k,m_k(\boldsymbol{x})}}, \forall k \in \mathcal{K}$, $\tau_1^*(\boldsymbol{x})$ can be obtained via the bisection search based on the equality $\sum_{k=1}^{K} \frac{\ln 2 I_k}{\tau_1^*(\boldsymbol{x})\left(-\mathcal{W}(-a_k e^{-a_k}) - a_k\right)} = B$, and $\mathcal{W}(x)$ denotes the Lambert function with $x = \mathcal{W}(x) e^{\mathcal{W}(x)}$ [14].

From (21), it is observed that $\bar{b}_k^*(\boldsymbol{x})$ is monotonically increasing w.r.t. I_k and decreasing w.r.t. $\frac{\bar{P}_k \bar{g}_{k,m_k(\boldsymbol{x})}}{N_0}$. By substituting (21) into (19), $\tau_1^*(\boldsymbol{x})$ can be obtained via the bisection search, and $\bar{\boldsymbol{b}}^*(\boldsymbol{x})$ can be directly obtained from (21). The complexity of bisection search for $\tau_1^*(\boldsymbol{x})$ is $\mathcal{O}(\log_2 T)$, and T is the initially chosen maximum value of τ_1.

B. Optimal Computation Allocation Solution to (P3)

Next, we consider problem (P3). By introducing a new variable $\lambda_2 = \frac{1}{\tau_2}$, problem (P3) is equivalently transformed into

$$\begin{aligned} \text{(P5)}: \max_{\{\tilde{f}_k \geq 0\}, \lambda_2 \geq 0} \quad & \lambda_2 \\ \text{s.t.} \quad & \sum_{k=1}^{K} x_{k,m} \tilde{f}_k \leq F_m, \forall m \in \mathcal{M} && (22) \\ & I_k W_k \lambda_2 \leq \tilde{f}_k, \forall k \in \mathcal{K}. && (23) \end{aligned}$$

It is easy to show that at the optimal solution to problem (P5), it must hold that $I_k W_k \lambda_2 = \tilde{f}_k, \forall k \in \mathcal{K}$. By substituting it into (22), we get a set of inequalities:

$$\lambda_2 \leq \frac{F_m}{\sum_{k=1}^{K} x_{k,m} I_k W_k}, \forall m \in \left\{ m' \in \mathcal{M} : \sum_{k=1}^{K} x_{k,m'} > 0 \right\}.$$

Therefore, we have the optimal solution to problem (P3) as

$$\tau_2^*(\boldsymbol{x}) = \frac{1}{\lambda_2^*(\boldsymbol{x})}, \tag{24}$$

$$\lambda_2^*(\boldsymbol{x}) = \min_{m \in \mathcal{M}: \sum_{k=1}^{K} x_{k,m} > 0} \frac{F_m}{\sum_{k=1}^{K} x_{k,m} I_k W_k}, \tag{25}$$

$$\tilde{f}_k^*(\boldsymbol{x}) = I_k W_k \lambda_2^*(\boldsymbol{x}), \forall k \in \mathcal{K}. \tag{26}$$

C. Optimal Downlink Bandwidth and Power Allocation to (P4)

Then, we consider problem (P4). Notice that $g_m(\underline{p}_l) \triangleq \log_2\left(1+\underline{p}_l \frac{\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{N_0}\right)$ is a concave function w.r.t. $\underline{p}_l$, and $\underline{b}_l \log_2\left(1+\frac{\underline{p}_l \underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\underline{b}_l N_0}\right) = \underline{b}_l g_m\left(\frac{\underline{p}_l}{\underline{b}_l}\right)$ is the perspective function of $g_m(\cdot)$. It thus follows that function $\underline{b}_l \log_2\left(1+\frac{\underline{p}_l \underline{g}_{m,l}}{\underline{b}_l N_0}\right)$ is jointly concave w.r.t. $\underline{b}_l$ and $\underline{p}_l$ [13]. Therefore, problem (P4) is convex, and thus can be optimally solved via the Lagrangian duality method [13]. Let $\beta \geq 0$ denote the dual variable associated with constraint (16), $\gamma_m \geq 0$ denote that associated with the m-th constraint in (17), $m \in \mathcal{M}$, and $\mu_l \geq 0$ denote that associated with the l-th constraint in (18), $l \in \mathcal{L}$. Accordingly, the partial Lagrangian of problem (P4) is expressed as

$$\begin{aligned} & \mathcal{L}\left(\{\underline{b}_l\}, \{\underline{p}_l\}, \tau_3, \{\mu_l\}, \beta, \{\gamma_m\}\right) \\ & = \tau_3 + \sum_{l=1}^{L} \mu_l \left(\frac{O_{\alpha(l)}}{\tau_3} - \underline{b}_l \log_2\left(1+\frac{\underline{p}_l \underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\underline{b}_l N_0}\right)\right) \\ & + \beta\left(\sum_{l=1}^{L} \underline{b}_l - B\right) + \sum_{l=1}^{L} \gamma_{m_{\alpha(l)}(\boldsymbol{x})} \underline{p}_l - \sum_{m=1}^{M} \gamma_m \underline{P}_m. \end{aligned} \tag{27}$$

Accordingly, the dual function of problem (P4) is given by

$$\text{(P6)}: f\left(\{\mu_l\},\beta,\{\gamma_m\}\right) \\ = \min_{\{0\leq \underline{b}_l \leq B\},\{\underline{p}_l\geq 0\},\tau_3\geq 0} \mathcal{L}\left(\{\underline{b}_l\},\{\underline{p}_l\},\tau_3,\{\mu_l\},\beta,\{\gamma_m\}\right).$$

We have the following lemma.

Lemma 1. In order for $f\left(\{\mu_l\},\beta,\{\gamma_m\}\right)$ to be bounded from below, for each $l\in\mathcal{L}$, it must hold that $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}>0$, or $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}=\mu_l=0$.

Proof. See the proof in [15]. □

The dual problem is defined as

$$\text{(P7)}: \max_{\{\mu_l\geq 0\},\beta\geq 0,\{\gamma_m\geq 0\}} f\left(\{\mu_l\},\beta,\{\gamma_m\}\right). \tag{28}$$

Since problem (P4) satisfies the Slater's condition, the strong duality holds between (P4) and its dual problem (P7). To solve (P4), we equivalently solve (P7) by first solving (P6) to obtain $f\left(\{\mu_l\},\beta,\{\gamma_m\}\right)$ with a given set of $\{\mu_l\}$, β, and $\{\gamma_m\}$, and then searching over $\{\mu_l\geq 0\}$, $\beta\geq 0$, and $\{\gamma_m\geq 0\}$ to maximize $f\left(\{\mu_l\},\beta,\{\gamma_m\}\right)$ in (28).

In particular, first, given a set of $\{\mu_l\geq 0\}$, $\beta\geq 0$, and $\{\gamma_m\geq 0\}$, problem (P6) can be decomposed into the following $L+1$ subproblems:

$$\min_{\tau_3\geq 0} \quad \tau_3+\frac{\sum_{l=1}^{L}\mu_l O_{\alpha(l)}}{\tau_3} \tag{29}$$

$$\min_{0\leq \underline{b}_l\leq B,\underline{p}_l\geq 0} -\mu_l\underline{b}_l\log_2\left(1+\frac{\underline{p}_l\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\underline{b}_l N_0}\right)+\beta\underline{b}_l \\ +\gamma_{m_{\alpha(l)}(\boldsymbol{x})}\underline{p}_l, \forall l\in\mathcal{L}. \tag{30}$$

For subproblem (29), it can be easily verified that the optimal solution is given by

$$\tau_3^{(\{\mu_l\},\beta,\{\gamma_m\})}=\sqrt{\sum_{l=1}^{L}\mu_l O_{\alpha(l)}}. \tag{31}$$

For each subproblem l in (30), it is easy to verify that if $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}=\mu_l=\beta=0$, then any $\underline{b}_l\geq 0$ and $\underline{p}_l\geq 0$ are the optimal solution; if $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}=\mu_l=0$ and $\beta>0$, $\underline{b}_l=0$ and any $\underline{p}_l\geq 0$ are optimal; and if $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}>0$ and $\mu_l=0$, then the optimal solution is $\underline{p}_l=0$. Since the optimal solutions under the above cases are trivial, we focus on the case when $\gamma_{m_{\alpha(l)}(\boldsymbol{x})}>0$ and $\mu_l>0$, and the optimal solutions to the L subproblems in (30) are given in the following lemma.

Lemma 2. Given $\{\mu_l>0\}$, $\beta\geq 0$, and $\{\gamma_m>0\}$, the optimal solution of $\{\underline{b}_l\}$ and $\{\underline{p}_l\}$ to the subproblems in (30) satisfies

$$\frac{\underline{p}_l^{(\{\mu_l\},\beta,\{\gamma_m\})}}{\underline{b}_l^{(\{\mu_l\},\beta,\{\gamma_m\})}}=\left\lceil\frac{\mu_l\log_2 e}{\gamma_{m_{\alpha(l)}(\boldsymbol{x})}}-\frac{N_o}{\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}\right\rceil^{\dagger}, \forall l\in\mathcal{L}, \tag{32}$$

$$\underline{b}_l^{(\{\mu_l\},\beta,\{\gamma_m\})}\begin{cases}=0, & \text{if } \beta>\epsilon_l,\\ \in[0,B], & \text{if } \beta=\epsilon_l,\\ =B, & \text{if } \beta<\epsilon_l,\end{cases} \quad \forall l\in\mathcal{L}, \tag{33}$$

where $\epsilon_l \triangleq \mu_l\left\lceil\log_2\left(\frac{\mu_l\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\ln 2N_0\gamma_{m_{\alpha(l)}(\boldsymbol{x})}}\right)\right\rceil^{\dagger} - \left\lceil\mu_l\log_2 e-\frac{\gamma_{m_{\alpha(l)}(\boldsymbol{x})}N_0}{\underline{g}_{m_{\alpha(l)}(\boldsymbol{x})}}\right\rceil^{\dagger}$, $l\in\mathcal{L}$, and $\lceil x\rceil^{\dagger}\triangleq\max\{x,0\}$.

Proof. See the proof in [15]. □

Next, with $\tau_3^{(\{\mu_l\},\beta,\{\gamma_m\})}$, $\left\{\underline{b}_l^{(\{\mu_l\},\beta,\{\gamma_m\})}\right\}$, and $\left\{\underline{p}_l^{(\{\mu_l\},\beta,\{\gamma_m\})}\right\}$ obtained, we consider problem (P7). Since problem (P7) is convex but not necessarily differentiable, the ellipsoid method [16] is adopted to solve it. The subgradients of $f\left(\{\mu_l\},\beta,\{\gamma_m\}\right)$ w.r.t. $\{\mu_l\}$, β and $\{\gamma_m\}$ are given as $\frac{O_{\alpha(l)}}{\tau_3}-\underline{b}_l\log_2\left(1+\frac{\underline{p}_l\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\underline{b}_l N_0}\right)$, $\sum_{l=1}^{L}\underline{b}_l-B$, and $\sum_{l=1}^{L}x_{\alpha(l),m}\underline{p}_l-\underline{P}_m$, respectively.

Finally, with the optimal dual solution $\{\mu_l^*\}$, β^*, and $\{\gamma_m^*\}$ at hand, $\tau_3^{(\{\mu_l^*\},\beta^*,\{\gamma_m^*\})}$ in (31) is the optimal solution $\tau_3^*(\boldsymbol{x})$ to problem (P4). It is easy to show that at the optimal solution, the following condition must hold:

$$\frac{O_{\alpha(l)}}{\tau_3}=\underline{b}_l\log_2\left(1+\frac{\underline{p}_l\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\underline{b}_l N_0}\right), \forall l\in\mathcal{L}. \tag{34}$$

By combining (32) and (34), $\underline{\boldsymbol{b}}^*(\boldsymbol{x})$ is given by

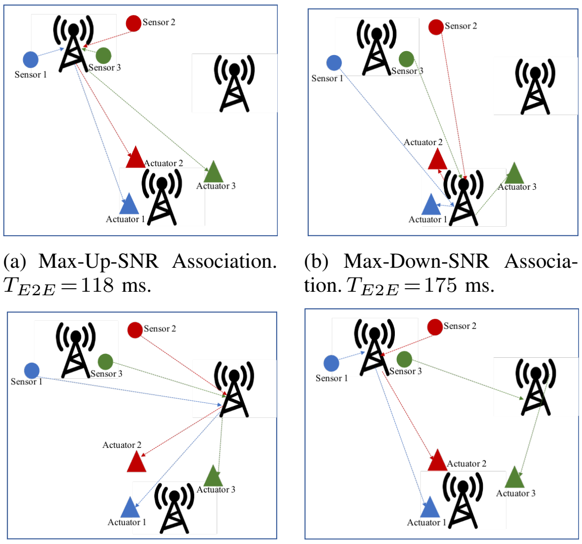
$$\underline{b}_l^*(\boldsymbol{x})=\frac{O_{\alpha(l)}}{\tau_3^*(\boldsymbol{x})\log_2\left(\frac{\mu_l^*\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}{\ln 2N_0\gamma^*_{m_{\alpha(l)}(\boldsymbol{x})}}\right)}, \forall l\in\mathcal{L} \tag{35}$$

and then $\underline{p}_l^*(\boldsymbol{x})=\left(\frac{\mu_l\log_2 e}{\gamma_{m_{\alpha(l)}(\boldsymbol{x})}}-\frac{N_o}{\underline{g}_{m_{\alpha(l)}(\boldsymbol{x}),l}}\right)\underline{b}_l^*(\boldsymbol{x}), \forall l\in\mathcal{L}$.

By combining the solutions in Sections III-A, III-B, and III-C, together with the brute-force search over $\boldsymbol{x}$, the optimal solution to problem (P1) is finally found.

IV. Numerical Results

In this section, we provide numerical examples to validate our results. Unless otherwise mentioned, we set $M=K=L=3$, $I_k=10$ Kbits, $O_k=2I_k$, $W_k=800$ cycles/bit, $N_0=-174$ dBm/Hz, $\bar{P}_k=\underline{P}_m=30$ dBm, $f_m=10$ GHz, $B=1$ KHz, $\alpha(l)=\text{mod}(L-l+1,K)+1$, $\forall k\in\mathcal{K},m\in\mathcal{M},l\in\mathcal{L}$. The path loss between any two nodes is modeled as $\beta_0\left(\frac{d}{d_0}\right)^{-\zeta}$, where $\beta_0=-30$ dB denotes the path loss at the reference distance $d_0=10$ m, d denotes the distance between them, and $\zeta=3$ denotes the path loss exponent. The proposed optimal algorithm is compared with the following three traditional heuristic baselines: *Max-Up-SNR*: Each sensor is associated with the AP that provides the maximal SNR for the uplink transmission, i.e., $x_{k,m^*}=1$ with $m^*\triangleq\arg\max_{m\in\mathcal{M}}\bar{g}_{k,m}$, and $x_{k,m}=0$, otherwise, $k\in\mathcal{K},m\in\mathcal{M}$. *Max-Down-SNR*: Each sensor is associated with the AP that provides the maximal sum of the SNR for the downlink transmission between the AP and the actuators that request the sensor, i.e., $x_{k,m^*}=1$, where $m^*\triangleq\arg\max_{m\in\mathcal{M}}\sum_{l\in\mathcal{L}_k}\underline{g}_{m,l}$, and $x_{k,m}=0$, otherwise, $k\in\mathcal{K},m\in\mathcal{M}$. *Max-Comp. Freq.*: Each sensor is associated with the MEC server that provides the maximal computation



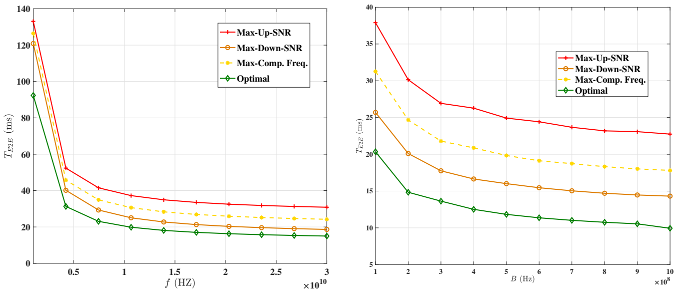
(a) Max-Up-SNR Association. $T_{E2E}=118$ ms.

(b) Max-Down-SNR Association. $T_{E2E}=175$ ms.

(c) Max-Comp. Freq. Association. $T_{E2E}=103$ ms.

(d) Optimal Association. $T_{E2E}=97$ ms.

Fig. 3: Simulation setup. $I_k=10$ Kbits, $O_k=20$ Kbits, $\bar{d}_{k,m}=3m$ km, $W_k=800$ cycles/bit, $\underline{d}_{m,l}=7-0.3m$ km, $f_1=10$ GHz, $f_2=30$ GHz and $f_3=20$ GHz, $k\in\mathcal{K}$, $m\in\mathcal{M}$, $l\in\mathcal{L}$.



(a) E2E latency versus frequency.

(b) E2E latency versus bandwidth.

Fig. 4: Performance comparison.

frequency, i.e., $x_{k,m^*}=1$, where $m^*\triangleq\arg\max_{m\in\mathcal{M}} f_m$, and $x_{k,m}=0$, otherwise, $k\in\mathcal{K}, m\in\mathcal{M}$. Based on the association policy, the resource allocation policy of each baseline is obtained via solving problem (P2), problem (P3) as well as problem (P4), respectively.

First, take Fig. 3 as a toy example. Assume that each actuator requests the input data generated from the sensor in the same color. We can see that heuristic Max-Up-SNR, Max-Down-SNR and Max-Comp. Freq. association policies incur larger latency and load imbalance than the optimal policy. Then, Fig. 4(a) and Fig. 4(b) compare the E2E latency obtained by the brute force search-based optimal solution with that by the three baselines versus different computation frequency values and bandwidth values, respectively. From Fig. 4, we see that the E2E latency of our proposed optimal algorithm achieves good performance gains over the three baselines.

V. Conclusion

This paper investigated a novel MEC-enabled IoT network with joint communication and computation load balancing and resource allocations to minimize the E2E latency of the sensing-communication-computation-actuation service loop. The formulated E2E latency minimization problem, however, is a mixed-integer non-linear program that is challenging to solve. To tackle this issue, we proposed the optimal solution by using the brute-force search. Numerical results were finally provided to show the performance gains achieved by our proposed design over conventional benchmarks without joint design. It is our hope that this paper can provide new design insights on MEC-enabled IoT networks with sensing-communication-computation-actuation loops.

References

[1] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the internet of things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[2] E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 110–117, Jun. 2017.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, Jan. 2017.

[4] K. Guo and T. Q. S. Quek, "On the asynchrony of computation offloading in multi-user MEC systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7746–7761, Dec. 2020.

[5] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.

[6] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7573–7586, Nov. 2019.

[7] J. Yan, S. Bi, Y. J. Zhang, and M. Tao, "Optimal task offloading and resource allocation in mobile-edge computing with inter-user task dependency," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 235–250, Jan. 2020.

[8] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.

[9] K. Guo, R. Gao, W. Xia, and T. Q. S. Quek, "Online learning based computation offloading in MEC systems with communication and computation dynamics," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1147–1162, Feb. 2021.

[10] K. Li, M. Tao, and Z. Chen, "Exploiting computation replication for mobile edge computing: A fundamental computation-communication tradeoff study," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4563–4578, Jul. 2020.

[11] L. Huang, X. Feng, L. Zhang, L. Qian, and Y. Wu, "Multi-server multi-user multi-task computation offloading for mobile edge computing networks," *Sensors*, vol. 19, no. 6, pp. 1–19, Mar. 2019.

[12] Q. Yuan, H. Zhou, J. Li, Z. Liu, F. Yang, and X. S. Shen, "Toward efficient content delivery for automated driving services: An edge computing solution," *IEEE Network*, vol. 32, no. 1, pp. 80–86, Jan./Feb. 2018.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Uni. Press, Mar. 2004.

[14] D. Barry, J.-Y. Parlange, L. Li, H. Prommer, C. Cunningham, and F. Stagnitti, "Analytical approximations for real values of the Lambert W-function," *Math. Comput. Simul.*, vol. 53, no. 1-2, pp. 95–103, Jul. 2000.

[15] Y. Sun, J. Xu, and S. Cui, "Joint user association and resource allocation optimization for MEC-Enabled IoT networks," Nov. 2021. [Online]. Available: https://drive.google.com/file/d/118LcdXF91peTekDCaqvNb_zlcJo-E7a_/view?usp=sharing.

[16] R. G. Bland, D. Goldfarb, and M. J. Todd, "The ellipsoid method: A survey," *Oper. Res.*, vol. 29, no. 6, pp. 1039–1091, Jul. 1981.