Delay Minimization for Massive MIMO Assisted Mobile Edge Computing

Ming Zeng, Wanming Hao, Octavia A. Dobre, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Mobile edge computing (MEC) has been envisioned as a promising technology for enhancing the computational capacities of mobile devices, by enabling computational task offloading. In this article, we employ massive multiple-input multiple-output methods to facilitate offloading in MEC. Our objective is to minimize the maximum delay for offloading and computing among the users, which requires a joint allocation of wireless and computational resources. Both perfect and imperfect channel state information (CSI) are considered. Under perfect CSI, we derive a semi-closed-form solution for the formulated problem. Under imperfect CSI, since the formulated problem is non-convex, we transform it into a convex one using a successive convex approximation technique and propose an iterative algorithm to solve it. Presented numerical results show the benefits of having a large number of antennas at the base station, and the necessity of performing joint radio and computational resource allocation.**

***Index Terms*—Massive multiple-input multiple-output (MIMO), mobile edge computing (MEC), delay minimization, joint resource allocation.**

I. Introduction

Mobile edge computing (MEC) has been recognized as a promising technology for 5G and future wireless communication networks, e.g., fiber-wireless networks [1], [2], internet-of-things networks [3], [4] and unmanned aerial vehicle aided communications [5]. MEC enhances the computation capacities and prolongs the lifespan of mobile devices, by enabling computational task offloading [6], [7]. In MEC systems, the energy consumption of the mobile devices and the overall delay for offloading and computing are deemed as two critical system metrics. Generally, these two metrics are in opposition to one another. In order to strike a balance between them, a joint allocation of radio and computational resources is often required.

Resource allocation in MEC is often a non-trivial problem, since two different resources, i.e., wireless and computational resources, need to be handled together. For the sake of tractability, most works on MEC focus on the simple scenario when each of the offloading users and the base station (BS) is equipped with a single antenna [8]–[12]. This, however, fails to exploit the advantages brought by multiple-input multiple-output (MIMO) technology in terms of offloading efficiency. To address this, a few works have considered MEC under the more practical scenario when the BS is equipped with multiple antennas [13], [14]. Specifically, the authors in [13] consider a single-cell MIMO system with perfect and imperfect channel state information (CSI), respectively, and the formulated weighted energy minimization problem is addressed using alternating optimization and difference of convex functions (DC) programming. The authors in [14] study a multi-cell MIMO system with a common edge server, and the formulated energy minimization problem is handled via a successive convex approximation technique.

Recently, research on massive MIMO enabled MEC has attracted significant attention [15]–[19]. By employing massive MIMO, more users can offload simultaneously [17]. Moreover, the offloading delay and energy consumption are reduced due to the increased spectral and energy efficiencies [18], [19]. More specifically, the authors in [16] consider a cell-free multi-cell massive MIMO system, and the impact of computation probability on the total energy consumption is analyzed using stochastic geometry and queuing theory. The authors in [17] apply millimeter wave and massive MIMO to standard wireless local area networks with MEC. A novel medium access control protocol is proposed by exploiting the unique characteristics of millimeter wave and massive MIMO systems. The authors in [18] study the delay minimization problem for a single cell massive MIMO enabled MEC network. Two heuristic algorithms are proposed to obtain suboptimal solutions. In addition, the authors in [19] consider the energy minimization problem for a massive MIMO enabled heterogeneous network with MEC. A low-complexity algorithm based on alternating optimization is proposed to jointly optimize the wireless and computing resources.

Considering user fairness, we aim to minimize the maximum delay for offloading and computing among the users under maximum transmit power and energy consumption constraints. Moreover, both perfect and imperfect CSI cases are considered. We derive a semi-closed-form solution for perfect CSI. However, the formulated problem is non-convex under the imperfect CSI case, and thus, is difficult to solve directly. By using the successive convex approximation technique, we transform the original problem into a convex one and further propose an iterative algorithm to solve it. Simulation results show the superiority of the proposed joint radio and computational resource allocation schemes over the baselines when computational resource allocation is fixed.

II. System Model and Problem Formulation

A. System Model

We consider a MEC system, where a BS equipped with a massive antenna array supports multiple single antenna users for computation offloading. Denote the set of users by $\mathcal{K} = \{1, \ldots, K\}$, and antennas by $\mathcal{M} = \{1, \ldots, M\}$. Each user $k \in \mathcal{K}$ generates computationally intensive tasks, characterized by two parameters: the size of the input data, L_k, and the number of CPU cycles required for processing, W_k. There exist two subsequent phases: offloading and computing. For the offloading phase, the users send their input data to the BS via the wireless channels. Upon receiving the data, the BS needs to allocate its computing power to the tasks and perform the computing for the users.

1) Communication Model: Two scenarios are considered: 1) perfect CSI at the BS, and 2) imperfect CSI at the BS. For both scenarios, it is assumed that the zero-forcing technique is adopted at the BS to suppress the inter-user interference.

Manuscript received July 28, 2019; revised January 20, 2020; accepted March 1, 2020. Date of publication March 9, 2020; date of current version June 18, 2020. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, through its Discovery Program and in part by the U.S. National Science Foundation under Grants CCF-0939370, CCF-1513915, and CCF-1908308. The review of this article was coordinated by Dr. J. Liu. *(Corresponding author: Octavia Dobre.)*

Ming Zeng is with the Memorial University, St. John's, NL A1B 3X9, Canada, and also with the Faculty of Science and Engineering, Laval University, Quebec, QC, G1V0A6, Canada (e-mail: mzeng@mun.ca).

Wanming Hao is with the Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450001, China (e-mail: wmhao@hotmail.com).

Octavia A. Dobre is with the Memorial University, St. John's, NL A1B 3X9, Canada (e-mail: iewmhao@zzu.edu.cn).

H. Vincent Poor is with the Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/TVT.2020.2979434

For the perfect CSI case, the zero-forcing detection matrix is given by $\mathbf{V} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$, where $\mathbf{H} \in \mathcal{C}^{M\times K}$ denotes the overall channel matrix. On this basis, we can obtain the normalized effective channel gain for user k as $h_k = \frac{|\mathbf{V}(k)\mathbf{H}_k|^2}{\sigma^2|\mathbf{V}(k)|^2}$, where $\mathbf{V}(k)$ is the k-th row of $\mathbf{V}$, $\mathbf{H}_k$ is the k-th column of $\mathbf{H}$, and σ^2 denotes the noise power. Accordingly, the achievable offloading data rate at user k is given by

$$R_k = \log_2\left(1 + P_k h_k\right), \tag{1}$$

where P_k is the corresponding transmit power, satisfying $P_k \leq P_k^{\max}$, with $P_k^{\max}$ being the maximum transmit power.

For the imperfect CSI case, we consider the practical scenario when CSI is obtained through uplink training. For simplicity, we assume that each user transmits at maximum power for training [20]. Denote the large-scale fading coefficient by β_k, which is known at the BS. Then, the mean-square of the channel estimate is given by $\gamma_k = \frac{\tau_{\mathrm{p}}\beta_k^2 P_k^{\max}/\sigma^2}{1+\tau_{\mathrm{p}}\beta_k P_k^{\max}/\sigma^2}, \forall k \in \mathcal{K}$, where $\tau_{\mathrm{p}} \geq K$ is the length of the pilot sequences. As a result, the achievable offloading data rate at user k can be approximated by [20]

$$R_k = \log_2\left(1 + \frac{P_k h_k}{1 + \sum_{i=1}^{K} P_i h_i^{'}}\right), \tag{2}$$

where P_k is the corresponding transmit power, while h_k and $h_i^{'}$ denote the normalized effective channel gains, satisfying $h_k = (M - K)\gamma_k/\sigma^2$ and $h_i^{'} = (\beta_i - \gamma_i)/\sigma^2$.

Accordingly, the transmission time for user k can be expressed as

$$T_k = \frac{L_k}{R_k}. \tag{3}$$

2) Computing Model: Let us denote the computational capacity of the MEC server by F, which is shared among all users. The computing resource allocated to user k is denoted by f_k, satisfying $\sum_{i=1}^{K} f_k = F$. Then, the computation time of user k's task is given by

$$Q_k = \frac{W_k}{f_k}. \tag{4}$$

Accordingly, the overall delay for offloading and computing at user k is $T_k + Q_k$.

B. Problem Formulation

Considering user fairness, we aim to minimize the maximum overall delay for offloading and computing among the users under maximum transmit power and energy consumption constraints. This requires a joint allocation of users‘ transmit power and the MEC server's computing resource. The considered delay minimization problem can be formulated as follows:

$$\mathrm{P1}: \min_{P_k, f_k} \max_{\forall k\in\mathcal{K}} T_k + Q_k \tag{5a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} f_k = F, \tag{5b}$$

$$P_k \leq P_k^{\max}, \quad \forall k \in \mathcal{K} \tag{5c}$$

$$P_k T_k \leq E_k^{\max}, \quad \forall k \in \mathcal{K} \tag{5d}$$

where $E_k^{\max}$ denotes the maximum energy constraint for user k. (5b) is the computing resource constraint. (5c) denotes the maximum power constraint, while (5d) represents the maximum energy constraint. Note that the time and energy consumption for channel estimation are neglected, since both are fixed, and do not affect the way of solving the problem.

III. Joint Resource Allocation for the Perfect CSI Case

To handle the min-max operation in P1, we introduce an auxiliary variable t, and reformulate P1 as

$$\mathrm{P2}: \min_{P_k, f_k, t} t \tag{6a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} f_k = F, \tag{6b}$$

$$P_k \leq P_k^{\max}, \quad \forall k \in \mathcal{K} \tag{6c}$$

$$\frac{P_k L_k}{\log_2\left(1 + P_k h_k\right)} \leq E_k^{\max}, \quad \forall k \in \mathcal{K} \tag{6d}$$

$$\frac{L_k}{\log_2\left(1 + P_k h_k\right)} + \frac{W_k}{f_k} \leq t, \quad \forall k \in \mathcal{K} \tag{6e}$$

Next, we re-write (6d) as $P_k L_k - E_k^{\max}\log_2(1 + P_k h_k) \leq 0$. On this basis, it is clear that P2 is a convex optimization problem, and thus, the optimal solution can be obtained using standard convex optimization techniques, e.g., interior-point method [21].

However, our interest is to find an analytical solution for further insight. According to (6e), t decreases with P_k. Therefore, to minimize t, a larger value of P_k is preferred. P_k is constrained by (6c) and (6d). Let us first look at (6d), and it can be verified that $\frac{P_k L_k}{\log_2(1+P_k h_k)}$ grows with P_k. Therefore, maximum P_k is obtained when equality is achieved at (6d), i.e., $P_k L_k = E_k^{\max}\log_2(1 + P_k h_k)$. After some algebraic manipulations, the root is given by

$$\hat{P}_k = -\frac{W\left(-\frac{L_k \ln 2}{E_k^{\max} h_k} 2^{-\frac{L_k}{E_k^{\max} h_k}}\right) E_k^{\max}}{L_k \ln 2} - \frac{1}{h_k}, \tag{7}$$

where $W(x)$ denotes the Lambert-W function, which is the inverse function of $f(z) = z\exp(z) = x$, i.e., $z = W(x)$. Note that $W(x)$ is a built-in function in most well-known mathematical software, e.g., Matlab. Combining (6c) and (6d), we can obtain the optimal P_k as $P_k^\star = \min(\hat{P}_k, P_k^{\max})$.

Now let us consider the allocation of f_k, and we have the following theorem:

Theorem 1: The maximum overall delay is minimized when the allocation of f_k satisfies $t = \frac{L_k}{\log_2(1+P_k^\star h_k)} + \frac{W_k}{f_k}, \forall k \in \mathcal{K}$, i.e., the overall delay for each user is the same.

Proof: The proof can be constructed using contradiction. Denote the optimal maximum overall delay by t. Assume that there exists at least one user whose overall delay is less than t. Then, we can easily shift the computing resource from such a user to all other users whose overall delay equal to t, such that t decreases. This contradicts our original assumption that t is optimal. Therefore, no user can achieve a lower overall delay than others. That is, all users achieve the same overall delay. This completes the proof. ∎

Based on Theorem 1, we have

$$f_k = \frac{W_k}{t - \frac{L_k}{\log_2\left(1+P_k^\star h_k\right)}}, \quad \forall k \in \mathcal{K} \tag{8}$$

It can be seen that $f_k, \forall k \in \mathcal{K}$ decreases when t increases. Meanwhile, $\sum_{k\in\mathcal{K}} f_k = F$ needs to hold. As a result, t can be obtained

Algorithm 1: Bisection Method for Computing Resource Allocation.

```
**Initialization:** $t_{\text{low}} \leftarrow \max(\frac{L_k}{\log_2(1+P_k^\star h_k)})$;
    $t_{\text{up}} \leftarrow \max(\frac{L_k}{\log_2(1+P_k^\star h_k)} + \frac{KW_k}{F})$; $\epsilon \leftarrow 10^{-6}$
**while** $t_{\text{up}} - t_{\text{low}} > \epsilon$
    $t_{\text{mid}} \leftarrow \frac{t_{\text{low}}+t_{\text{up}}}{2}$;
    $f_k \leftarrow \frac{W_k}{t_{\text{mid}} - \frac{L_k}{\log_2(1+P_k^\star h_k)}}, \quad \forall k \in \mathcal{K}$;
    **if** $\sum_{k\in\mathcal{K}} f_k < F$
        $t_{\text{up}} \leftarrow t_{\text{mid}}$;
    **else**
        $t_{\text{low}} \leftarrow t_{\text{mid}}$;
    **end**;
**end**;
```

using the bisection method.[1] For initialization, the lower and upper bound of t are given by $t_{\text{low}} = \max(\frac{L_k}{\log_2(1+P_k^\star h_k)})$ and $t_{\text{up}} = \max(\frac{L_k}{\log_2(1+P_k^\star h_k)} + \frac{KW_k}{F})$, respectively. The specific procedure is given in Algorithm 1. Once t is obtained, we can attain f_k directly from (8).

Complexity analysis: The proposed solution consists of two steps: 1) calculating $P_k^\star$ based on the Lambert-W function; and 2) calculating f_k using Algorithm 1. The Lambert-W function is often solved via the Newton's method, and has a quadratic convergence rate. Denote the iterations for obtaining $P_k^\star$ by I_1, which is often a small number. Then, the complexity of step 1) is $O(KI_1)$. The bisection method has a linear convergence rate, and the number of iterations for obtaining t is $\log_2(\frac{t_{\text{up}}-t_{\text{low}}}{\epsilon})$. We need to calculate f_k for all users in Algorithm 1, and thus, the complexity for step 2) is $O(K\log_2(\frac{t_{\text{up}}-t_{\text{low}}}{\epsilon}))$. As a result, the overall complexity is $O(K(I_1 + \log_2(\frac{t_{\text{up}}-t_{\text{low}}}{\epsilon})))$, which is a linear function of the number of users K.

IV. Joint Resource Allocation for the Imperfect CSI Case

Likewise, we introduce an auxiliary variable t to handle the min-max operation in P1, and reformulate P1 as

$$\text{P3}: \min_{P_k, f_k, t} \; t \tag{9a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} f_k = F, \tag{9b}$$

$$P_k \le P_k^{\max}, \quad \forall k \in \mathcal{K} \tag{9c}$$

$$\frac{P_k L_k}{\log_2\left(1 + \frac{P_k h_k}{1+\sum_{i=1}^K P_i h_i'}\right)} \le E_k^{\max}, \quad \forall k \in \mathcal{K} \tag{9d}$$

$$\frac{L_k}{\log_2\left(1 + \frac{P_k h_k}{1+\sum_{i=1}^K P_i h_i'}\right)} + \frac{W_k}{f_k} \le t, \quad \forall k \in \mathcal{K} \tag{9e}$$

Problem P3 is non-convex due to the non-convex constraints (9d) and (9e). Nonetheless, we still have the following theorem:

Theorem 2: The overall delay for each user under the optimal solution is the same, i.e., $t = \frac{L_k}{\log_2(1+\frac{P_k h_k}{1+\sum_{i=1}^K P_i h_i'})} + \frac{W_k}{f_k}, \forall k \in \mathcal{K}$.

Proof: The proof follows the same structure as Theorem 1, and is omitted. ∎

Although Theorem 2 provides a condition that the optimal solution needs to satisfy, it is still not enough to solve (9). To address this, we next transform the non-convex constraints (9d) and (9e) into convex ones by advanced convex approximation techniques.

Specifically, we first introduce an auxiliary variable w_k, satisfying

$$w_k \le \frac{P_k h_k}{1+\sum_{i=1}^K P_i h_i'}, \quad \forall k \in \mathcal{K}. \tag{10}$$

On this basis, (9d) and (9e) can be re-expressed as

$$\frac{P_k L_k}{\log_2(1+w_k)} \le E_k^{\max} \Longleftrightarrow P_k L_k \le \log_2(1+w_k) E_k^{\max}, \quad \forall k \in \mathcal{K} \tag{11a}$$

$$\frac{L_k}{\log_2(1+w_k)} + \frac{W_k}{f_k} \le t, \quad \forall k \in \mathcal{K}. \tag{11b}$$

It can be easily verified that both (11a) and (11b) are convex.

To handle (10), we introduce another auxiliary variable y as follows

$$w_k y \le P_k h_k, \quad \forall k \in \mathcal{K}, \tag{12}$$

where

$$y \ge 1 + \sum_{k=1}^{K} P_k h_i'. \tag{13}$$

The upper bound of $w_k y$ is given by [22]

$$\frac{y^{[n]}}{2w_k^{[n]}} w_k^2 + \frac{w_k^{[n]}}{2y^{[n]}} y^2 \ge w_k y, \tag{14}$$

where $y^{[n]}$ and $w_k^{[n]}$ denote the values of y and w_k at the n-th iteration, respectively. Then, (12) can be reformulated as the following convex constraint

$$\frac{y^{[n]}}{2w_k^{[n]}} w_k^2 + \frac{w_k^{[n]}}{2y^{[n]}} y^2 \le P_k h_k. \tag{15}$$

Now we can reformulate P3 as

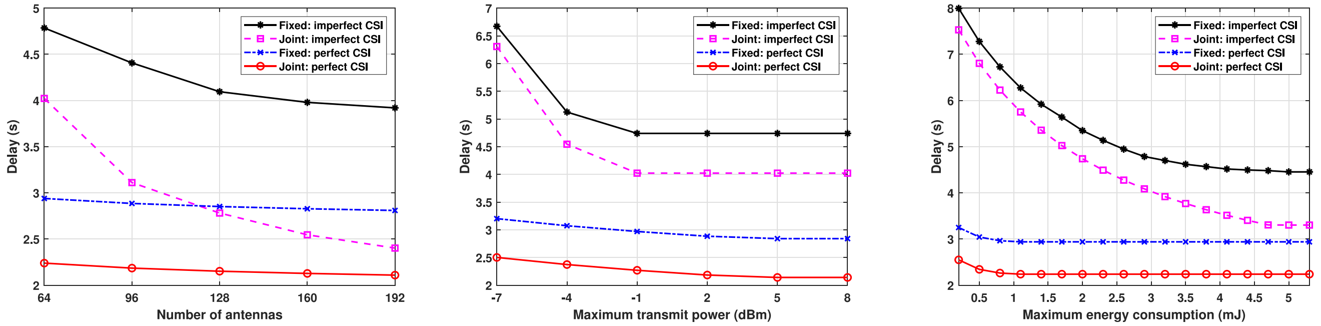
$$\text{P4}: \; \min t \text{ s.t. } (9b), (9c), (11a), (11b), (13),(15). \tag{16}$$

It is clear that P4 is a convex optimization problem, and can be solved by standard convex optimization technique, e.g., interior-point method [21]. Note that the only approximation procedure adopted in transforming P3 to P4 is (15). Because of (15), we need to iteratively solve P4 to obtain the solution of P3. Specifically, starting from an initial feasible solution, we update $y^{[n]}$ and $w_k^{[n]}$ iteratively by solving P4 using the obtained results from the previous iteration. The above procedure is carried out until convergence.

Convergence proof: At each iteration, we solve the convex optimization problem P4. The obtained optimal solution will yield a decrease or at least equal value of t. Since the objective function clearly has a lower bound, e.g., $t = 0$, convergence is guaranteed.

Complexity analysis: The proposed solution requires to solve P4 iteratively. The computational complexity of solving P4 is $O([3K+2]^{3.5})$ [23], where $3K+2$ denotes the number of variables. Accordingly, the overall computational complexity of the proposed solution is $O(I_2[3K+2]^{3.5})$, where I_2 denotes the number of iterations for convergence.

[1] t can also be obtained by finding the root of $\sum_{k=1}^{K} \frac{W_k}{t - \frac{L_k}{\log_2(1+P_k^\star h_k)}} - F = 0$ using the Newton's method.



(a) (b) (c)

Fig. 1. Delay as a function of (a) antenna number, (b) maximum transmit power, and (c) maximum energy consumption.

V. Numerical Results

In this section, numerical results are presented to evaluate the performance of the proposed schemes. The default simulation parameters are as follows: there are $K = 8$ users, which are uniformly generated within a radius of 300 m. The pathloss model follows $30.6 + 36.7 \log_{10}(d)$, where d is the distance in m. Rayleigh fading is used for small-scale fading. The bandwidth is $B = 1$ MHz, while the noise power spectral density is $N_0 = -174$ dBm/Hz. The antenna number is $M = 64$, while the total CPU computing capacity at the MEC server is $F = 20$ G cycles/s. The maximum transmit power is $P_k^{\max} = 0$ dBm, while the maximum energy consumption is $E_k^{\max} = 2$ mJ. The data length is $L_k = 10$ Mbits, whereas the computing need for user $k, k \in \{1, \ldots, K\}$ is $W_k = k/2 + 1$ G cycles.

For both perfect and imperfect CSI cases, we consider the corresponding special case with equal computing resource allocation among the users as the baseline scheme. The power allocation is then obtained by using the proposed solutions with $f_k, \forall k \in \mathcal{K}$ fixed to F/K.

Fig. 1(a)-(c) show how the maximum delay varies with the antenna number, maximum transmit power and maximum energy consumption, respectively. Here "Joint" denotes the proposed scheme, while "Fixed" represents the baseline algorithm with equal computing resource among the users. According to Fig. 1(a), the maximum delay declines with the antenna number for all considered algorithms. The reasons are twofold: 1) users' offloading data rates increase with the number of antennas under both perfect and imperfect CSI, and thus, less time is needed for offloading [20]; 2) a reduced offloading time further yields a lower overall delay for both the proposed scheme and the baseline algorithm. The benefit of having more antennas at the BS is fully illustrated from Fig. 1(a). Meanwhile, the proposed scheme outperforms the baseline algorithm under both perfect and imperfect CSI. This is because the proposed scheme jointly optimizes the radio and computational resources, while the baseline algorithm only optimizes the transmit power under even computational resource allocation. The necessity of conducting joint radio and computational resources allocation is verified here. In particular, the proposed scheme under imperfect CSI can achieve lower delay than the baseline algorithm under perfect CSI when $M \geq 128$. As can be seen from Fig. 1(b), the delay first decreases with the maximum transmit power, and then remains fixed for all algorithms. The initial decrease in delay is owing to the increased offloading rates resulting from the increased transmit power. However, when $P_k^{\max}$ is large enough, the delay is constrained by the energy consumption rather than $P_k^{\max}$. Further increase in $P_k^{\max}$ brings no benefit, and thus, the delay remains fixed. Meanwhile, perfect CSI achieves lower delay than imperfect CSI for both schemes, because inter-user interference is eliminated under perfect CSI. As for Fig. 1(c), a similar trend can be observed for all algorithms as in Fig. 1(b), i.e., the delay first decreases with the energy consumption, and then remains fixed. This is because the delay is constrained by the maximum transmit power instead of the maximum energy consumption when $E_k^{\max}$ is large enough.

VI. Conclusion

In this article, we have considered the overall delay minimization among all users for a massive MIMO assisted MEC system under maximum transmit power and energy consumption constraints. The formulated problem requires a joint radio and computational resource allocation. For the perfect CSI case, we have derived a semi-closed-form solution, and further proposed an optimal and low-complexity algorithm. For the imperfect CSI case, the formulated non-convex problem has been handled by the successive convex approximation technique. Simulation results show the obtained performance gain when employing more antennas at the BS, especially under imperfect CSI. Moreover, it is clear that the proposed joint resource allocation schemes significantly outperform the baseline ones with fixed computing resource allocation. Finally, it can be seen that the maximum transmit power and energy consumption constraints are coupled and both need to be increased to fully lower the delay.

References

[1] H. Guo and J. Liu, "Collaborative computation offloading for multiaccess edge computing over fiber-wireless networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4514–4526, May 2018.

[2] H. Guo, J. Zhang, and J. Liu, "FiWi-enhanced vehicular edge computing networks: Collaborative task offloading," *IEEE Veh. Technol. Mag.*, vol. 14, no. 1, pp. 45–53, Mar. 2019.

[3] H. Guo, J. Zhang, J. Liu, and H. Zhang, "Energy-aware computation offloading and transmit power allocation in ultradense IoT networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4317–4329, Jun. 2019.

[4] H. Guo, J. Liu, J. Zhang, W. Sun, and N. Kato, "Mobile-edge computation offloading for ultradense IoT networks," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4977–4988, Dec. 2018.

[5] J. Xiong, H. Guo, and J. Liu, "Task offloading in UAV-aided edge computing: Bit allocation and trajectory optimization," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 538–541, Mar. 2019.

[6] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 3, pp. 1628–1656, Jul.–Sep. 2017.

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

[8] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

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

[10] 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 Trans. Commun.*, vol. 65, no. 8, pp. 3571–3584, Aug. 2017.

[11] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint computation offloading and interference management in wireless cellular networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432–7445, Aug. 2017.

[12] 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.

[13] T. T. Nguyen, L. Le, and Q. Le-Trung, "Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation," *IEEE Trans. Services Comput.*, p. 1, 2019.

[14] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[15] M. Zeng, W. Hao, O. A. Dobre, Z. Ding, and H. V. Poor, "Massive MIMO-assisted mobile edge computing: Exciting possibilities for computation offloading," *IEEE Veh. Technol. Mag.*, vol. 15, no. 2, pp. 31–38, Jun. 2020.

[16] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, p. 1, 2020.

[17] Y. Zhao, X. Xu, Y. Su, L. Huang, X. Du, and N. Guizani, "Multi-user MAC protocol for WLANs in MmWave massive MIMO systems with mobile edge computing," *IEEE Access*, vol. 7, pp. 181 242–181 256, 2019.

[18] T. Huang *et al.*, "Joint pilot and data transmission power control and computing resource allocation for the massive MIMO based MEC network," in *Proc. IEEE Int. Conf. Commun. Technol.*, Oct. 2019, pp. 860–865.

[19] Y. Hao, Q. Ni, H. Li, and S. Hou, "Energy-efficient multi-user mobile-edge computation offloading in massive MIMO enabled HetNets," in *Proc. IEEE Int. Conf. Commun.*, May 2019, pp. 1–6.

[20] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[22] P. Song, G. Scutari, F. Facchinei, and L. Lampariello, "D3M: Distributed multi-cell multigroup multicasting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 3741–3745.

[23] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 804–808, Jul. 2015.