Dynamic Spectrum Sharing for Load Balancing in Multi-Cell Mobile Edge Computing

Ming Zeng, *Student Member*, *IEEE*, and Viktoria Fodor, *Member*, *IEEE*

Abstract—**Large-scale mobile edge computing (MEC) systems require scalable solutions to allocate communication and computing resources to the users. In this letter we address this challenge by applying dynamic spectrum sharing among the base stations (BSs), together with local resource allocation in the cells. We show that the network-wide resource allocation can be transformed into a convex optimization problem, and propose a distributed, hierarchical solution with limited information exchange among the BSs. Numerical results demonstrate that the proposed solution is superior to other baseline algorithms, when wireless and computing resource allocation is not jointly optimized, or the wireless resources allocated to the BSs are fixed.**

Index Terms—**MEC, multi-cell, resource allocation**

I. INTRODUCTION

By enabling mobile devices to offload computation-intensive tasks to servers in close proximity, mobile edge computing (MEC) can provide low-latency services for emerging applications, such as immersive augmented reality, wearable cognitive assistance, or autonomous driving. Meanwhile, computation offloading can decrease the energy consumption of the mobile devices [1] and thus prolong their lifetime.

Early works on MEC focus on single cell systems with multiple users [1]–[3]. Recently, the general scenario of multi-cell MEC is receiving attention [4]–[7]. In [4], a MIMO multicell system with a common edge server is considered. The formulated energy minimization problem is solved using successive convex approximation. A game theoretic approach for the joint optimization of wireless and computing resources is proposed in [5], while the performance of MEC in heterogeneous networks is evaluated in [6], using stochastic geometry. A comprehensive study on the complexity of service placement and request routing in multi-cell MEC is provided in [7]. Most of the above works consider resource allocation in the multi-cell MEC as a large, centralized optimization problem, an approach that is not viable for large-scale systems. Research on cellular networks faced the same issue, and provided the approaches of biasing (also called cell breathing) [8], [9], and dynamic spectrum sharing (also called channel borrowing) [10]–[12] to balance network traffic across the cells. Initial results for biasing in MEC are shown in [6].

In this letter we adapt dynamic spectrum sharing to achieve communication and computation load balancing among the BSs, with the objective to minimize the total transmission

energy consumption under computational delay constraints [1]. We show that energy minimization can be transformed into a convex optimization problem, for which centralized optimal solution exists. Based on the centralized problem formulation, we propose a primal-dual resource allocation algorithm that lends itself to an iterative distributed solution, where BSs cooperate to share the spectrum, while each individual BS allocates its local communication and computing resources to the associated users. Numerical results show that the joint resource allocation can reduce the energy consumption significantly, while the proposed distributed solution requires limited information exchange among the BSs and converges within a few iterations.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a MEC system that consists of K users, and M BSs, each equipped with a MEC server. The users offload their computation tasks to a BS for processing. We denote the set of users by $\mathcal{K} = \{1, \cdots, K\}$, and the set of BSs by $\mathcal{M} = \{1, \cdots, M\}$. We consider that each user $i \in \mathcal{K}$ generates computationally intensive and delay sensitive tasks, characterized by three parameters, the size L_i of the input data, the number W_i of CPU cycles required to perform the computation, and the completion time constraint D_i.

The objective of the considered MEC system is to minimize the energy consumption for data transmission under the delay constraint, by jointly allocating the wireless and computing resources, as well as the transmission power of the users.

Communication resources: The overall system bandwidth is B Hz. We consider flat fading channel and orthogonal access with frequency division multiple access. Users are associated to the BS with the best received signal-to-noise ratio, as it is often the case in today's cellular systems. Denote the corresponding channel gain for user i by h_i. Then, the achievable data rate at user i is given by $R_i = x_i \log_2 \left(1 + \frac{P_i h_i}{x_i N_0} \right)$, where P_i is the corresponding transmission power, and x_i denotes the allocated bandwidth, satisfying $\sum_{i \in \mathcal{K}} x_i = B$. Besides, N_0 is the noise power spectral density coefficient. Accordingly, the transmission time and the resulting transmission energy consumption are respectively given by $T_i = \frac{L_i}{R_i}$ and $E_i = \frac{L_i P_i}{R_i}$.

Note that we consider orthogonal spectrum access here to reveal insights on joint resource allocation in MEC. Extension to multi-cell MEC systems with frequency reuse is discussed in Section V.

Computing resources: Let us denote the computational capacity of the MEC server at BS $j, j \in \mathcal{M}$ by C_j and the set of users associated with BS j by $\mathcal{S}_j, |\mathcal{S}_j| = K_j$. The

users served by the BS j, i.e., $\forall i \in \mathcal{S}_j$ share the computing resource of the MEC server. We denote the computing resource allocated to user i as q_i, satisfying $\sum_{i \in \mathcal{S}_j} q_i = C_j$. Then, the computational time of user i's task is given by $Q_i = \frac{W_i}{q_i}$ [13].

Energy consumption minimization: We consider the problem of total transmission energy minimization, under the constraint on the completion time of the computational tasks. That is, for each user i, the sum of the transmission and computational times should not violate the maximum delay D_i, i.e., $T_i + Q_i \leq D_i$. The delay constraint then can be turned into the following rate requirement: $R_i \geq \frac{L_i}{D_i - Q_i}$.

The energy minimization problem can be formulated as

$$P1 : \min_{\mathbf{P}, \mathbf{x}, \mathbf{q}} \sum_{i \in \mathcal{K}} E_i \tag{1a}$$

$$\text{s.t.} \quad R_i \geq \frac{L_i}{D_i - Q_i}, \forall i \in \mathcal{K} \tag{1b}$$

$$\sum_{i \in \mathcal{K}} x_i = B \tag{1c}$$

$$\sum_{i \in \mathcal{S}_j} q_i = C_j, \forall j \in \mathcal{M} \tag{1d}$$

where $\mathbf{P} \in \mathbb{R}^K, \mathbf{x} \in \mathbb{R}^K, \mathbf{q} \in \mathbb{R}^K$ are the vectors of allocated powers P_i, bandwidth x_i and computational resource q_i, respectively. Inequality constraints (1b) reflect the minimum data rate requirement for each user. Constraints (1c) limit the bandwidth, while (1d) restrict the computing resource.

III. CENTRALIZED RESOURCE ALLOCATION

To solve P1, the wireless and computing resources need to be allocated jointly. They are however coupled in a non-linear way through the delay constraint. To progress with the solution, we first state the following theorem.

Theorem 1: Under any given bandwidth and computing resource allocation $\mathbf{x}, \mathbf{q}$, the energy consumption is minimized when $T_i + Q_i = D_i, \forall i \in \mathcal{K}$ holds and the transmission power is set as $P_i = \frac{N_0 x_i}{h_i} \left(2^{\frac{R_i^{\min}}{x_i}} - 1 \right), \forall i \in \mathcal{K}$ where $R_i^{\min}$ is the minimum rate that still fulfills the delay requirement, i.e., $R_i^{\min} = \frac{L_i}{D_i - Q_i}$.

Proof: When x_i and q_i are given, the energy consumption of the users is independent, and minimizing the total energy consumption is equivalent to minimizing that of each user. Without loss of generality, we look at E_i, which can be reformulated as $E_i = \frac{L_i P_i}{R_i} = \frac{L_i P_i}{x_i \log_2 \left(1 + \frac{P_i h_i}{x_i N_0} \right)}$. Clearly, E_i increases with P_i, and therefore, E_i is minimized when the minimum power is used. Meanwhile, to satisfy the delay constraint, we have $R_i = x_i \log_2 \left(1 + \frac{P_i h_i}{x_i N_0} \right) \geq R_i^{\min}$, i.e., $P_i \geq (2^{R_i^{\min}/x_i} - 1) N_0 x_i / h_i$. At equality the achieved rate is $R_i^{\min}$, which in turn results a transmission time of $T_i = D_i - Q_i$. This concludes the proof. ∎

Let us then reformulate P1, based on Theorem 1. In addition, let us replace variables q_i with

$$t_i = D_i - W_i / q_i. \tag{2}$$

This then leads to

$$P2 : \min_{\mathbf{x}, \mathbf{t}} \sum_{i \in \mathcal{K}} \frac{N_0}{h_i} x_i t_i \left(2^{\frac{L_i}{x_i t_i}} - 1 \right) \tag{3a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{K}} x_i = B \tag{3b}$$

$$\sum_{i \in \mathcal{S}_j} \frac{W_i}{D_i - t_i} = C_j, \forall j \in \mathcal{M} \tag{3c}$$

In P2, equality (3c) is clearly not affine, and thus, the feasible set is non-convex. To address it, we relax the equality constraint and substitute (3c) with

$$\sum_{i \in \mathcal{S}_j} \frac{W_i}{D_i - t_i} \leq C_j, \forall j \in \mathcal{M} \tag{4}$$

As a consequence of Theorem 1, for any user i, the energy consumption decreases if q_i, the computing resource allocated to the user is increased. Thus, for the optimal solution, equality is achieved in (4), which means substituting (3c) with (4) will not change the solution.

Theorem 2: Problem P2 with the relaxed constraint (4) is a convex optimization problem.

Proof: First, equality constraint (3b) is affine. Then, for inequality constraint (4), its second derivative is $\sum_{i \in \mathcal{S}_j} \frac{2W_i}{(D_i - t_i)} > 0$, and thus, it is convex. Last, let us consider the objective function (3a). It can be seen that the energy consumption for each user is only affected by its own variables, e.g., for user i, $N_0 x_i t_i \left(2^{\frac{L_i}{x_i t_i}} - 1 \right) / h_i$ is only affected by x_i and t_i. Therefore, we can consider each user separately. Without loss of generality, we consider user i, whose Hessian matrix is given by

$$\mathbf{H}_i = \frac{N_0}{h_i} \cdot \begin{bmatrix} \mathbf{H}_i(1,1) & \mathbf{H}_i(1,2) \\ \mathbf{H}_i(2,1) & \mathbf{H}_i(2,2) \end{bmatrix},$$

where $\mathbf{H}_i(1,1) = \ln 2^2 \cdot 2^{\frac{L_i}{t_i x_i}} \cdot \frac{L_i^2}{t_i x_i^3}$, while $\mathbf{H}_i(2,2) = \ln 2^2 \cdot 2^{\frac{L_i}{x_i t_i}} \cdot \frac{L_i^2}{x_i t_i^3}$. Besides, $\mathbf{H}_i(1,2) = \mathbf{H}_i(2,1) = 2^{\frac{L_i}{x_i t_i}} - 1 + \ln 2^2 \cdot \frac{L_i^2}{x_i^2 t_i^2} 2^{\frac{L_i}{x_i t_i}} - \ln 2 \cdot \frac{L_i}{x_i t_i} 2^{\frac{L_i}{x_i t_i}}$. After some algebraic manipulations, it can be verified that $\det(\mathbf{H}_i) > 0$ holds for all $\frac{L_i}{x_i t_i} > 0$, which indicates (3a) is convex. This completes the proof. ∎

Based on Theorem 2, the optimal solution of P2 can be obtained using standard convex optimization methods in a centralized manner.

IV. DISTRIBUTED RESOURCE ALLOCATION WITH DYNAMIC SPECTRUM SHARING

In this section we propose an Iterative Resource Allocation algorithm to solve problem P2, that lends itself to a distributed implementation, with decreased signaling needs. As shown in Algorithm 1, it follows two iterative steps: i) the Bandwidth Allocation Algorithm (BAA) updates $\mathbf{x}$ to allocate bandwidth across and within the BSs, for given $\mathbf{t}$, and ii) the Computation resource Allocation Algorithm (CAA) updates $\mathbf{t}$ to allocate the computing resource at each BS, for given bandwidth allocation $\mathbf{x}$. We denote by E_i^t and E_i^x the energy consumption of user i after optimizing t_i and x_i, respectively, and ϵ is the stop condition.

Algorithm 1 Iterative Resource Allocation

1: **Initialization:** $q_i \leftarrow C_j/K_j, t_i \leftarrow \left(D_i - \frac{W_i}{q_i}\right), \forall i \in \mathcal{S}_j, j \in \mathcal{M}$;

2: Update $x_i, \forall i \in \mathcal{K}$ based on BAA, and calculate $\sum_i E_i^x$;

3: $\sum_i E_i^t \leftarrow \sum_i E_i^x + 2\epsilon$;

4: **while** $\sum_i E_i^t - \sum_i E_i^x > \epsilon$ **do**

5:       Update t_i based on CAA, and recalculate $\sum_i E_i^t$;

6:       Update x_i based on BAA, and recalculate $\sum_i E_i^x$;

7: **end while**

The Bandwidth Allocation Algorithm (BAA): Assuming that the computing resource allocation $\mathbf{t}$ is given, problem P2 is simplified as

$$\text{P3}: \quad \min_{\mathbf{x}} \sum_{i \in \mathcal{K}} \frac{N_0}{h_i} t_i x_i \left(2^{\frac{L_i}{t_i x_i}} - 1\right) \text{ s.t. (3b).} \quad (5a)$$

Since $\mathbf{H}_i(1,1) > 0$, P3 is a convex problem, and we can use the Karush-Kuhn-Tucker (KKT) condition to derive the optimal $\mathbf{x}$. The KKT condition for user i is

$$g(x_i) = \frac{N_0 t_i}{h_i} \left[2^{\frac{L_i}{t_i x_i}} - \frac{L_i}{t_i x_i} 2^{\frac{L_i}{t_i x_i}} \ln 2 - 1\right] + \lambda = 0, \forall i \in \mathcal{K}$$

where λ is the introduced auxiliary variable, satisfying $\lambda > 0$. For given λ, the above equation can be used to obtain x_i. Specifically, we have $\frac{\partial g(x_i)}{\partial x_i} = \frac{\ln 2^2 \cdot N_0 L_i^2}{h_i t_i x_i^3} 2^{\frac{L_i}{t_i x_i}} > 0$, which indicates that $g(x_i)$ grows with x_i, and thus a bisection search can be used to obtain x_i by comparing $g(x_i)$ with 0. Now the problem lies in how to obtain λ. When λ is increased, $x_i, \forall i \in \mathcal{K}$ will decrease to ensure $g(x_i) = 0$. Meanwhile, $\sum_i x_i = B$ needs to hold. Consequently, λ can also be obtained with bisection search, by comparing $\sum_i x_i$ with B.

The resulting BAA consists of two loops: an outer loop to find the value of λ and an inner loop to determine the bandwidth allocation $\mathbf{x}$.

The Computing resource Allocation Algorithm (CAA): Under given bandwidth allocation, the computing resource allocation is independent across the BSs. Thus, the energy minimization for each BS is equivalent to that of the overall system. Let us consider BS j and user set $\mathcal{S}_j$, $j \in \mathcal{M}$. The corresponding optimization problem can be formulated as

$$\text{P4}: \quad \min_{\mathbf{t}} \sum_{i \in \mathcal{S}_j} \frac{N_0}{h_i} x_i t_i \left(2^{\frac{L_i}{x_i t_i}} - 1\right) \text{ s.t. (4).} \quad (6)$$

As P3, P4 is also a convex problem, and the KKT condition is given by

$$f(t_i) = \frac{N_0 x_i}{h_i} \left[2^{\frac{L_i}{x_i t_i}} - \frac{L_i}{x_i t_i} 2^{\frac{L_i}{x_i t_i}} \ln 2 - 1\right] + \frac{W_i}{(D_i - t_i)^2} \mu_j = 0,$$
$$\forall i \in \mathcal{S}_j, j \in \mathcal{M}$$

where μ_j is the introduced auxiliary variable, satisfying $\mu_j \geq 0$.

Since $\frac{\partial f(t_i)}{\partial t_i} = \frac{\ln 2^2 \cdot N_0 L_i^2}{h_i x_i t_i^3} 2^{\frac{L_i}{x_i t_i}} + \frac{4 W_i}{(D_i - t_i)^2} \mu_j > 0$, we can conclude that $f(t_i)$ grows with t_i, and further t_i declines with μ_j. Therefore, t_i and μ_j can be found with bisection search. At each BS $j, j \in \mathcal{M}$, the resulting CAA includes an outer loop to find the value of μ_j and an inner loop to determine $\mathbf{t}$, which in turn gives the computing resource allocation $\mathbf{q}$, according to (2).

Distributed Implementation with Dynamic Spectrum Sharing among the BSs:

The Iterative Resource Allocation algorithm requires the implementation of BAA and CAA, and the exchange of the parameters between these algorithms. Note that CAA can be performed by the individual BSs. Similarly, for BAA, the update of x_i under given λ can happen locally at the BS. Finding the appropriate λ value for the KTT condition however requires collaboration. Specifically, the BSs need to share their $\sum_{i \in \mathcal{S}_j} x_i$ values, that is, the bandwidth that should be allocated to BS j, and increase or decrease λ in the bisection search, if $\sum_{j \in \mathcal{M}} \sum_{i \in \mathcal{S}_j} x_i$ is larger or smaller than B.

Optimality and complexity:

Theorem 3: The Iterative Resource Allocation algorithm gives the optimal resource allocation in finite steps, with predefined accuracy ϵ.

Proof: In both lines 5 and 6 of Algorithm 1, the energy consumption decreases, or remains unchanged. Since there is a lower bound for the energy consumption, e.g., 0, the Iterative Resource Allocation algorithm always terminates, either by reaching the lower bound, or by achieving a decrease less than ϵ. Moreover, the obtained local optimum is also the global optimum since the considered problem is convex. ∎

The centralized implementation requires the collection of user parameters and the distribution of the resource allocation vectors to the BSs, thus, the signaling complexity is $\mathcal{O}(K)$, where K is the total number of users in the multi-cell system. The computational complexity comes form the iterations of Algorithm 1, where both BAA and CAA perform bisection search for λ and μ as well as for the x_i and t_i values. This gives a computational complexity of $\mathcal{O}(NK)$, where N is the number of iterations in Algorithm 1.

The distributed implementation requires information exchange among the BSs, to search for λ in BAA, in each iteration steps of Algorithm 1. This leads to a signaling overhead of $\mathcal{O}(NM)$, where M is the number of BSs. Each BS j needs to run BAA and CAA locally, and thus, the computation complexity is $\mathcal{O}(NK_j)$.

The distributed implementation has good scalability properties, however, the complexity depends on the number of iterations N. Therefore, in Section VI we investigate how N depends on the network parameters.

V. Multi-cell MEC with Frequency Reuse

The previous sections consider orthogonal spectrum allocation among cells to reduce the complexity of the analysis and reveal insights. Frequency reuse is however necessary in large systems to increase spectrum efficiency. To this end, let us first consider the case when fixed frequency reuse (i.e., according to 3 or 7 cell pattern) is adopted to avoid co-channel interference.
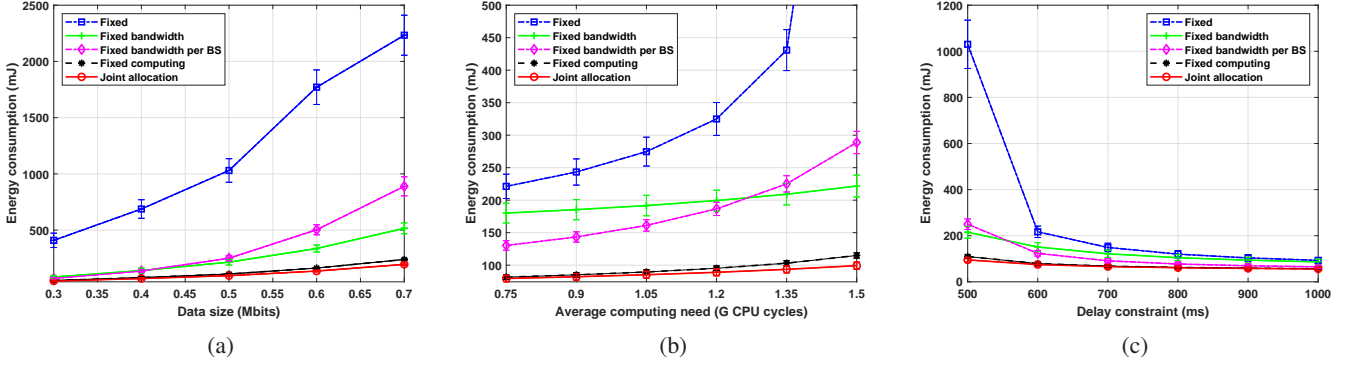
Fig. 1: Energy consumption as a function of (a) the data size, (b) the average computing need, and (c) the delay constraint.

In this case, we can reformulate P2 as

$$\text{P5}: \min_{\mathbf{x}, \mathbf{t}, B_f} \sum_{i \in \mathcal{K}} \frac{N_0}{h_i} x_i t_i \left(2^{\frac{L_i}{x_i t_i}} - 1 \right) \tag{7a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{S}_j} x_i = B_f, \forall j \in \mathcal{M}_f, f = \{1, \cdots, F\} \tag{7b}$$

$$\sum_{f=1}^{F} B_f = B \tag{7c}$$

$$\sum_{i \in \mathcal{S}_j} \frac{W_i}{D_i - t_i} = C_j, \forall j \in \mathcal{M} \tag{7d}$$

where F denotes the cell reuse factor, and $\mathcal{M}_f$ represents the cell set using the same frequency band B_f, $f = \{1, \cdots, F\}$. Then, (7b) denotes the bandwidth constraint for each cell, while (7c) is the total bandwidth constraint. Both (7b) and (7c) are affine constraints, and thus problem P5 with a relaxed (7d) (i.e., (4)) is convex, and can be easily solved using standard convex optimization tools.

An iterative solution that also allows distributed implementation can follow the lines of Algorithm 1. CAA can be performed as described in Section IV, but the bandwidth allocation algorithm has to be extended. Now the KKT conditions are given by

$$\frac{N_0 t_i}{h_i} \left[2^{\frac{L_i}{t_i x_i}} - \frac{L_i}{t_i x_i} 2^{\frac{L_i}{t_i x_i}} \ln 2 - 1 \right] + \lambda_j = 0, \forall i \in \mathcal{S}_j, j \in \mathcal{M} \tag{8}$$

$$\beta = \sum_{j \in \mathcal{M}_f} \lambda_j, \forall f = \{1, \cdots, F\} \tag{9}$$

where λ_j and β are the introduced auxiliary variables for (7b) and (7c), respectively, satisfying $\lambda_j, \beta > 0$.

Algorithm 2 summarizes the steps to find B_f, x_i, λ_j and β. The algorithm has an inner loop to determine B_f, x_i and λ_j for given β, according to BAA in Section IV and (9). This iteration ensures that the bandwidth is optimally allocated for given β values. Then, an outer loop finds β, such that constraint (7c) is satisfied. The distributed implementation requires β, B_f and λ_j to be exchanged among the cells.

Now let us consider the extreme case with universal frequency reuse. Due to the existence of co-channel interference, users' achievable rates are non-convex functions over

their transmit powers. As a result, the energy minimization problem is likely to be NP-hard [12, Theorem 1]. To make it tractable, we may need to refer to convex approximation or dual optimization [4], [12], [14].

Algorithm 2 Bandwidth Allocation with Frequency Reuse

1: **Initialization:** β_{low}; β_{up}; ϵ
2: **while** $\beta_{\text{up}} - \beta_{\text{low}} > \epsilon$
3:      $\beta \leftarrow \frac{\beta_{\text{low}} + \beta_{\text{up}}}{2}$;
4:      **for** $f \leftarrow 1, \cdots, F$
5:          initialization: B_f^{low}, B_f^{up};
6:          **while** $B_f^{\text{up}} - B_f^{\text{low}} > \epsilon$
7:              $B_f \leftarrow \frac{B_f^{\text{low}} + B_f^{\text{up}}}{2}$;
8:              obtain $x_i, \lambda_j, i \in \mathcal{S}_j, j \in \mathcal{M}_f$ as in BAA;
9:              **if** $\sum_{j \in \mathcal{M}_f} \lambda_j > \beta$ then $B_f^{\text{low}} \leftarrow B_f$;
10:             **else** $B_f^{\text{up}} \leftarrow B_f$;
11:             **end**;
12:          **end while**;
13:      **end for**;
14:      **if** $\sum_{f=1}^{F} B_f > B$ then $\beta_{\text{low}} \leftarrow \beta$
15:      **else** $\beta_{\text{up}} \leftarrow \beta$
16:      **end**;
17: **end while**

VI. NUMERICAL RESULTS

We evaluate the performance of the joint bandwidth and computing resource allocation scheme in a simulator implemented in Matlab. For each trial, we place the BSs and the users randomly uniformly in a disk of a radius of 200 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. We set $B = 10$ MHz, $N_0 = -174$ dBm/Hz and $\epsilon = 10^{-6}$.

We consider four baseline algorithms: a) equal bandwidth and computing resource per user, referred to as Fixed; b) equal bandwidth per user, while the computing resource is optimized, referred to as Fixed bandwidth; c) equal bandwidth for each BS, but optimized joint resource allocation within each BS, referred to as Fixed bandwidth per BS; and d) equal computing resource per user, with optimized bandwidth, referred to as Fixed computing.
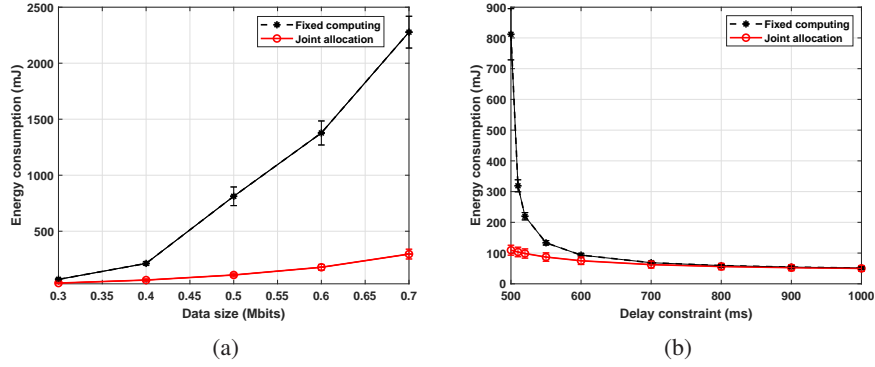
Fig. 2: Energy consumption for Fixed computing and Joint allocation as a function of (a) data size, and (b) delay constraint.

Fig. 1 shows the energy consumption under the five algorithms when the data size, the average computing need, and delay constraint vary, respectively. The default simulation values are: $M = 4$, $K = 32$, $C_j = 100$ G CPU cycles/s, $L_i = 0.5$ Mbits, $D_i = 500$ ms. For Figs 1(a) and (c), W_i is generated randomly uniformly within $[0.5, 2.5]$ G CPU cycles for each user. In Fig. 1(b), W is increased, and for each user W_i is generated randomly uniformly within $[\frac{1}{3}W, \frac{5}{3}W]$.

As expected, the energy consumption grows with the data size and computing need, but decreases as delay constraint gets relaxed. The proposed Joint resource allocation always achieves the best performance. The large difference between Joint allocation and Fixed bandwidth illustrates the gain of optimizing the bandwidth allocation among the users. Likewise, the difference between Joint allocation and Fixed bandwidth per BS indicates that the load in the cells can be highly unbalanced, and thus dynamic bandwidth sharing among the BSs is necessary. Fixed computing has similar performance to Joint allocation, the reason is probably that the disparity among users' computing needs is small in the considered scenario. Therefore, in Fig. 2 we present the corresponding results with a higher variance, i.e., W_i is generated randomly uniformly within $[0.5, 4]$ G CPU cycles for each user. It can be seen that Joint allocation consumes much lower energy than Fixed computing, especially under large data size or strict delay constraint.

We also conducted extensive simulations to evaluate N, the number of iterations required for the Iterative Resource Allocation algorithm to converge. We found that N does not depend significantly on M, the number of BSs, for example, for $M = 16$ and $K = 64$ the algorithm converges in two iterations on average. However, N increases almost linearly with K_j, the number of users in a cell. For example, under $M = 4$, the average number of iterations increases from two to four when K_j changes from $K_j = 8$ to $K_j = 16$.

VII. CONCLUSION

In this paper, we considered a multi-user multi-cell MEC system, where users offload their computing tasks to the BS with the best channel for processing. An overall transmission energy minimization problem was formulated and transformed into a convex optimization problem. Furthermore, a scalable

distributed solution was proposed inspired by the dynamic spectrum sharing approach in cellular networks. Numerical results showed that the proposed joint allocation outperforms other baseline algorithms, when wireless and computing resources are not jointly optimized, or the wireless resources allocated to the BSs are fixed.

REFERENCES

[1] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.
[2] 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.
[3] M. Zeng and V. Fodor, "Energy-efficient resource allocation for noma-assisted mobile edge computing," in *Proc. IEEE PIMRC*, Sep. 2018.
[4] 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.
[5] S. Josilo and G. Dan, "Joint allocation of computing and wireless resources to autonomous devices in mobile edge computing," in *Proc. ACM SIGCOMM Wkshps*, Aug. 2018.
[6] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks," *arXiv preprint arXiv:1804.07756,*, Apr. 2018.
[7] K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Joint service placement and request routing in multi-cell mobile edge computing networks," in *Proc. IEEE INFOCOM*, May 2019.
[8] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[9] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in hetnets: A utility perspective," *IEEE J. Select. Areas Commun.*, vol. 33, no. 6, pp. 1025–1039, Jun. 2015.
[10] I. Koutsopoulos and L. Tassiulas, "Joint optimal access point selection and channel assignment in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 521–532, June 2007.
[11] L. M. O. Khanbary and D. P. Vidyarthi, "A GA-based effective fault-tolerant model for channel allocation in mobile computing," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1823–1833, May 2008.
[12] Z. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Select. Areas Commun.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.
[13] Z. Liang, Y. Liu, T. Lok, and K. Huang, "Multiuser computation offloading and downloading for edge computing with virtualization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4298–4311, Sep. 2019.
[14] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.