Energy-Efficient Multi-User Mobile-Edge Computation Offloading in Massive MIMO Enabled HetNets

Yuanyuan Hao*, Qiang Ni+, Hai Li*, and Shujuan Hou*
*School of Information and Electronics, Beijing Institute of Technology, Beijing, China
+School of Computing and Communications, Lancaster University, Lancaster, U.K
Emails: {tracyhao, haili, shujuanhou}@bit.edu.cn, q.ni@lancaster.ac.uk

***Abstract*—In this paper, we investigate the energy-efficient multi-user mobile-edge computing offloading problem in massive MIMO enabled HetNets, where the CPU-cycle frequency of mobile devices, uplink power control, computational task offloading ratio and uplink transmission duration are jointly optimized. The problem is formulated as minimizing the energy consumption of all mobile devices while satisfying the maximum latency requirement. Specifically, to address this non-convex problem, a low-complexity algorithm is proposed relied on alternating optimization, where we address the joint computational task offloading ratio and uplink transmission duration optimization problem and the uplink power control problem iteratively. Besides, the effectiveness and convergence of the proposed iterative algorithm are analytically studied. Numerical results demonstrate that our proposed algorithm consumes less energy compared to local computing and full uploading schemes, and the application of massive MIMO in HetNets helps to reduce energy consumption of mobile devices.**

***Index Terms*—Computation offloading, energy efficiency, HetNets, massive MIMO, mobile edge computing.**

I. Introduction

With the striking growth of mobile computation-intensive applications, such as online gaming, limited battery energy and finite computation capacities become new challenges for mobile devices in the fifth generation (5G) networks. Mobile edge computing (MEC) is one promising solution which can offload intensive computation to nearby servers at the edge of cellular networks for remote execution [1], [2]. Meanwhile, as mobile data traffic demand is explosively increasing, massive multiple input multiple output (MIMO) and dense heterogeneous networks (HetNets) are introduced in 5G networks to enhance the system spectral efficiency (SE) and energy efficiency (EE) [3]–[6].

The studies about MEC systems have emerged recently [7]–[12]. In [7] and [8], the authors study joint radio and computational resource allocation for a single mobile user, where they formulate the problem as minimizing the mobile user's energy consumption and the task execution latency. Similarly, the work in [9] investigates the energy consumption minimization problem in multi-user MEC systems based on both time-division and orthogonal frequency-division multiple access. Besides, the authors in [10] provide insights on the tradeoff between power and delay in MEC systems, where an online computation offloading algorithm is designed relied on Lyapunov optimization. Nevertheless, the proposed algorithms in [9] and [10] are limited to interference-free scenarios.

The work in [11] studies the mobile-edge computation offloading problem in HetNets, and a centralized scheme is proposed to minimize the total energy consumption. A similar system is considered in [12], while a distributed scheme is designed to lower the system overhead in the field of energy and monetary cost. However, in [11] and [12], the MEC server is only located at the macro base station (MBS) and massive MIMO is not taken into account for uplink transmission. As shown in the recent studies [13]–[15], the combination of massive MIMO and HetNets can boost the system EE and SE. Nevertheless, as far as we know, the performances of MEC in massive MIMO enabled HetNets have not been evaluated in the existing literature, where the inter-cell interference should be taken into account.

Motivated by the above observations, this paper studies the energy-efficient multi-user computation offloading problem in massive MIMO enabled HetNets where each BS is equipped with an MEC server. To be specific, the main contributions of this paper are summarized in the following.

- Different from the existing researches [7]–[12], we consider the problem of minimizing the energy consumption of all mobile devices under the maximum latency requirement in massive MIMO enabled HetNets, which is the first time to jointly optimize the CPU clock speed of mobile devices, uplink power control, computational task offloading ratio and uplink transmission duration.
- In Propositions 1-2, we prove the effectiveness and convergence of the proposed algorithm theoretically. Simulation results also confirm that the proposed algorithm converges fast and significantly outperforms both local computing and full uploading schemes. Besides, it is shown that the combination of massive MIMO and HetNets in MEC benefits mobile energy savings.

II. System Model and Problem Formulation

Consider a multi-user MEC system with one MBS equipped with massive MIMO, $I-1$ single-antenna small cell BSs (SBS), and N single-antenna users in a set $\mathcal{C}=\{1,2,\cdots,N\}$. Let $i\in\mathcal{I}=\{1,2,\cdots,I\}$ denote the i-th BS, where the MBS corresponds to $i=1$, and the cases of $i>1$ represent SBSs. Each BS is equipped with one MEC server to carry

out computational tasks offloaded from its associated users. Besides, assume that the total frequency band is shared by all BSs. Each user can be only associated with one BS, which is determined prior to the resource allocation.

A. Uplink Data Rate Expression

Let p_n^{T} denote the transmit power of user n, and g_{in} represents the channel power gain between the i-th BS and the n-th user. With the adoption of the maximum-ratio combing detector, the lower bound of the achievable uplink data rate for users associated with the MBS is expressed as [16]

$$r_{in} = \log_2 \left(1 + \frac{(M-1)\, p_n^{\mathrm{T}} g_{in}}{\sum\limits_{m \neq n, m \in \mathcal{C}} p_m^{\mathrm{T}} g_{im} + \sigma_n^2} \right), i = 1, \quad (1)$$

where M denotes the number of antennas equipped at the MBS, and σ_n^2 means the noise power experienced by user n.

On the other hand, assume that users associated with the same SBS share the same time-frequency resource equally, and thus their uplink data rates can be calculated by

$$r_{in} = \frac{1}{N_i} \log_2 \left(1 + \frac{p_n^{\mathrm{T}} g_{in}}{\sum\limits_{m \notin \mathcal{C}_i} p_m^{\mathrm{T}} g_{im} + \sigma_n^2} \right), i > 1, \quad (2)$$

where $\mathcal{C}_i$ indicates the set of users associated with the i-th BS, and N_i is the cardinal number of $\mathcal{C}_i$.

Then, let $\mathbf{x} = [x_{in}]$ denote the user association matrix, where $x_{in} = 1$ if user n is linked to the i-th BS, and $x_{in} = 0$, otherwise. Thus, the uplink data rate of the n-th user is obtained as

$$R_n = \sum_{i \in \mathcal{I}} x_{in} r_{in}. \quad (3)$$

B. Partial Computing Offloading

Assume that a_n bits are required to be computed for user n in one time slot T. In this paper, we consider flexible data partition task model, and let s_n denote the ratio of offloaded data bits to the total input bits, i.e., $(1 - s_n)\, a_n$ for local computing, and $s_n a_n$ for edge computing.

For local computing, the power consumption for user n can be modelled as [2]

$$p_n^{\mathrm{L}} = \lambda f_n^3, \quad (4)$$

where f_n is the CPU-cycle frequency of user n and can be adjusted via dynamic voltage and frequency scaling (DVFS) technique [1]. Thus, the time for local computing of user n is calculated as

$$t_n^{\mathrm{L}} = \frac{\varepsilon\, (1 - s_n)\, a_n}{f_n}, \quad (5)$$

where ε ($\varepsilon > 0$) denotes the number of cycles needed per input data bit. Note that the parameter values a_n and ε are determined by the types of applications and estimated via task profilers [17]. Consequently, the energy consumption for local computing is given by

$$E_n^{\mathrm{L}} = p_n^{\mathrm{L}} t_n^{\mathrm{L}} = \lambda \varepsilon\, (1 - s_n)\, a_n f_n^2. \quad (6)$$

As for edge computing, users first offload tasks to their associated BSs. After collecting input bits from users, MEC servers at BSs execute offloaded tasks. Define T^{ul} as the specific duration for uplink data transmission, and the execution time for MEC servers is bounded by $T - T^{\mathrm{ul}}$. Note that the time for users to download computed results is considered to be negligible in this paper, as the results are usually of small size and BSs are with high transmit power. Thus, the energy consumption of user n for uplink transmission is expressed as

$$E_n^{\mathrm{T}} = p_n^{\mathrm{T}} T^{\mathrm{ul}}. \quad (7)$$

C. Problem Formulation

To minimize the energy consumption of N users while ensuring their tasks successfully executed in one time slot T, the energy-efficient multi-user computation offloading problem is formulated as

$$\min_{\mathbf{s}, \mathbf{p}^{\mathrm{T}}, \mathbf{f}, T^{\mathrm{ul}}} \sum_{n \in \mathcal{C}} \left(E_n^{\mathrm{L}} + E_n^{\mathrm{T}} \right) \quad (8a)$$

$$\text{s.t. } \mathrm{C1}: 0 \leq s_n \leq 1, \forall n, \quad (8b)$$

$$\mathrm{C2}: 0 \leq p_n^{\mathrm{T}} \leq p_{\max}^{\mathrm{T}}, \forall n, \quad (8c)$$

$$\mathrm{C3}: 0 \leq f_n \leq f_{\max}, \forall n, \quad (8d)$$

$$\mathrm{C4}: 0 \leq T^{\mathrm{ul}} \leq T, \quad (8e)$$

$$\mathrm{C5}: \varepsilon\, (1 - s_n)\, a_n / f_n \leq T, \forall n, \quad (8f)$$

$$\mathrm{C6}: s_n a_n - R_n T^{\mathrm{ul}} \leq 0, \forall n, \quad (8g)$$

$$\mathrm{C7}: \sum_{n \in \mathcal{C}_i} \varepsilon s_n a_n - f_i^{\mathrm{BS}} \left(T - T^{\mathrm{ul}} \right) \leq 0, \forall i, \quad (8h)$$

where $p_{\max}^{\mathrm{T}}$ and $f_{\max}$ are the maximum transmit power and frequency for each user, respectively, and f_i^{BS} denotes the CPU-cycle frequency of the MEC server at BS i.

In problem (8), C1 gives the range of computational task offloading ratio s_n; C2 and C3 represent that the maximum transmit power and CPU-cycle frequency of user n are $p_{\max}^{\mathrm{T}}$ and $f_{\max}$, respectively; C4 means that the time for uplink transmission T^{ul} should be not more than the time slot T; C5 indicates that the delay for local computing is bounded by T; C6 and C7 represent that the uplink data transmission and edge computing should be finished in the duration T^{ul} and $T - T^{\mathrm{ul}}$, respectively.

III. Joint Radio and Computational Resource Optimization

A. Optimal CPU Frequency and Problem Decomposition

Observing problem (8), it can be found that the objective function monotonously increases with f_n, $\forall n$. Besides, from constraint C5, we have $f_n \geq \frac{\varepsilon(1-s_n)a_n}{T}$. Thus, the optimal CPU-cycle frequency of user n can be obtained as

$$f_n^* = \frac{\varepsilon\, (1 - s_n)\, a_n}{T}, \quad (9)$$

on the condition that $\frac{\varepsilon(1-s_n)a_n}{T} \leq f_{\max}$.

Since a function can be always minimized by first minimizing it over some of the variables and then over the

remaining ones [18], by substituting (9) into (8), problem (8) is equivalently transformed into

$$\begin{aligned}&\min_{\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}} \Upsilon\left(\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}\right)=\sum_{n\in\mathcal{C}}\left(\frac{\lambda\varepsilon^3(1-s_n)^3a_n^3}{T^2}+p_n^{\mathrm{T}}T^{\mathrm{ul}}\right)\\&\text{s.t. C1, C2, C4, C6, C7,}\\&\quad \mathrm{C8}:\varepsilon\left(1-s_n\right)a_n-f_{\max}T\leq 0.\end{aligned}\tag{10}$$

Nevertheless, the transformed problem (10) is still non-convex, and finding its optimum is rather challenging. Hence, we divide it into two sub-problems, and solve them alteratively.

By fixing the transmit power vector $\mathbf{p}^{\mathrm{T}}$, the joint computational task offloading ratio and uplink transmission duration optimization problem can be obtained as

$$\begin{aligned}&\min_{\mathbf{s},T^{\mathrm{ul}}} \Upsilon\left(\mathbf{s},T^{\mathrm{ul}}\right)=\sum_{n\in\mathcal{C}}\left(\frac{\lambda\varepsilon^3(1-s_n)^3a_n^3}{T^2}+p_n^{\mathrm{T}}T^{\mathrm{ul}}\right)\\&\text{s.t. C1, C4, C6}-\mathrm{C8},\end{aligned}\tag{11}$$

which is convex and can be solved by standard algorithms, such as interior-point method with polynomial computational complexity required [18].

Conversely, fixing the computational task offloading ratio vector $\mathbf{s}$ and the uplink transmission duration T^{ul}, the uplink power control problem can be expressed as

$$\begin{aligned}&\min_{\mathbf{p}^{\mathrm{T}}} \Phi\left(\mathbf{p}^{\mathrm{T}}\right)=\sum_{n\in\mathcal{C}}p_n^{\mathrm{T}}T^{\mathrm{ul}},\\&\text{s.t. C2, C6.}\end{aligned}\tag{12}$$

B. Power Control Based on Sequential Optimization

Unfortunately, problem (12) is non-convex due to the existence of inter-cell interference shown in the user rate function R_n. Alternatively, R_n can be re-expressed as

$$R_n\left(\mathbf{p}^{\mathrm{T}}\right)=\sum_{i\in\mathcal{I}}x_{in}\left(a_{in}\left(\mathbf{p}^{\mathrm{T}}\right)-b_{in}\left(\mathbf{p}^{\mathrm{T}}\right)\right),\tag{13}$$

where $a_{in}\left(\mathbf{p}^{\mathrm{T}}\right)$ and $b_{in}\left(\mathbf{p}^{\mathrm{T}}\right)$ are defined as

$$a_{in}\left(\mathbf{p}^{\mathrm{T}}\right)=\begin{cases}\log_2\left(\left(M-1\right)p_n^{\mathrm{T}}g_{in}+\sum\limits_{m\neq n}p_m^{\mathrm{T}}g_{im}+\sigma_n^2\right), i=1,\\ \frac{1}{N_i}\log_2\left(p_n^{\mathrm{T}}g_{in}+\sum\limits_{m\notin\mathcal{C}_i}p_m^{\mathrm{T}}g_{im}+\sigma_n^2\right), i>1,\end{cases}\tag{14a}$$

$$b_{in}\left(\mathbf{p}^{\mathrm{T}}\right)=\begin{cases}\log_2\left(\sum\limits_{m\neq n}p_m^{\mathrm{T}}g_{im}+\sigma_n^2\right), i=1,\\ \frac{1}{N_i}\log_2\left(\sum\limits_{m\notin\mathcal{C}_i}p_m^{\mathrm{T}}g_{im}+\sigma_n^2\right), i>1,\end{cases}\tag{14b}$$

respectively.

It can be found from (13) and (14) that the rate function $R_n\left(\mathbf{p}^{\mathrm{T}}\right)$ is the difference of two concave functions. To address this power control problem with polynomial complexity, we adopt sequential optimization [19], which helps to obtain a series of improved solutions. To be specific, with an initial point $\mathbf{p}^{\mathrm{T},(0)}$, the rate function $R_n\left(\mathbf{p}^{\mathrm{T}}\right)$ is close to

$$\hat{R}_n^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)=\sum_{i\in\mathcal{I}}x_{in}\left(a_{in}\left(\mathbf{p}^{\mathrm{T}}\right)-\hat{b}_{in}^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)\right)\tag{15}$$

Algorithm 1 The power control algorithm based on sequential optimization

1. Initialize $t=0$, $\chi=1$, and a feasible point $\mathbf{p}^{\mathrm{T},(0)}$.
2. **while** $\chi>0.01$**, do**
3. $t=t+1$;
4. Calculate $\hat{b}_{in}^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)$ according to (16), $\forall i,n$;
5. Update constraint C6′;
6. Solve (17) and obtain its global optimum $\mathbf{p}^{\mathrm{T},(t)}$;
7. Calculate $\chi=\max\limits_{n}\left|\frac{p_n^{\mathrm{T},(t)}-p_n^{\mathrm{T},(t-1)}}{p_n^{\mathrm{T},(t-1)}}\right|$;
8. **end while**

at the t-th iteration where $\hat{b}_{in}^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)$ is the first-order approximation of $b_{in}\left(\mathbf{p}^{\mathrm{T}}\right)$, i.e.,

$$\begin{aligned}\hat{b}_{in}^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)&=b_{in}\left(\mathbf{p}^{\mathrm{T},(t-1)}\right)\\&\quad+\sum_{n\in\mathcal{C}}\left(p_n^{\mathrm{T}}-p_n^{\mathrm{T},(t-1)}\right)\left.\frac{\partial b_{in}\left(\mathbf{p}^{\mathrm{T}}\right)}{\partial p_n^{\mathrm{T}}}\right|_{\mathbf{p}^{\mathrm{T}}=\mathbf{p}^{\mathrm{T},(t-1)}}.\end{aligned}\tag{16}$$

Since $\hat{R}_n\left(\mathbf{p}^{\mathrm{T}}\right)$ is concave, the uplink power control solution of problem (12) can be found by solving the following convex optimization problem

$$\begin{aligned}&\min_{\mathbf{p}^{\mathrm{T}}} \tilde{\Phi}\left(\mathbf{p}^{\mathrm{T}}\right)=\sum_{n\in\mathcal{C}}p_n^{\mathrm{T}}T^{\mathrm{ul}},\\&\text{s.t. C2, C6}':\ s_na_n-\hat{R}_n^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)T^{\mathrm{ul}}\leq 0,\forall n.\end{aligned}\tag{17}$$

To tighten constraint C6′, $\hat{R}_n^{(t)}\left(\mathbf{p}^{\mathrm{T}}\right)$ should be updated iteratively, and problem (17) is updated correspondingly and solved until convergence. Specific steps are listed in **Algorithm 1**, whose *convergence* and *optimality* are demonstrated as follows.

Proposition 1: The objective function $\Phi\left(\mathbf{p}^{\mathrm{T}}\right)$ of problem (12) is decreased with the increasing iteration times of Algorithm 1, and the point of convergence satisfies the Karush-Kuhn-Tucher (KKT) conditions of problem (12).

Proof: Suppose that $\mathbf{p}^{\mathrm{T},(t)}$ is the obtained solution of Algorithm 1 at the t-th iteration, and it can be obtained that

$$\Phi\left(\mathbf{p}^{\mathrm{T},(t-1)}\right)\overset{(\mu)}{=}\tilde{\Phi}\left(\mathbf{p}^{\mathrm{T},(t-1)}\right)\overset{(\nu)}{\geq}\tilde{\Phi}\left(\mathbf{p}^{\mathrm{T},(t)}\right)\overset{(\omega)}{=}\Phi\left(\mathbf{p}^{\mathrm{T},(t)}\right).\tag{18}$$

Specifically, the equation (μ) is straightforward since problem (12) and (17) are equal at $\mathbf{p}^{\mathrm{T},(t-1)}$. The inequality (ν) holds because $\mathbf{p}^{\mathrm{T},(t)}$ is the optimal solution of problem (17). Since $\mathbf{p}^{\mathrm{T},(t)}$ satisfies constraint C6′ and C6′ is stricter than C6, $\mathbf{p}^{\mathrm{T},(t)}$ must satisfy C6, which validates the last equality (ω). Therefore, $\Phi\left(\mathbf{p}\right)$ is reduced with the increasing iteration times.

Since the constraint set of problem (12) is compact and $\Phi\left(\mathbf{p}\right)$ is lower-bounded, Algorithm 1 is guaranteed to converge. Suppose that $\mathbf{p}^{\mathrm{T}*}$ is the convergent point. Thus, $\mathbf{p}^{\mathrm{T}*}$ must satisfy the KKT conditions of problem (12), since problem (12) and problem (17) have the same objective and derivative values at $\mathbf{p}^{\mathrm{T}*}$. ■

C. Alternating Optimization Algorithm

With the above results, the alternating optimization algorithm for joint radio and computational resource optimization

Algorithm 2 Alternating optimization algorithm for joint radio and computational resource optimization

1. Initialize $k=0$, $\delta_s=\delta_T=1$, and a feasible point $\mathbf{p}^{\mathrm{T},(0)}$.
2. **while** $\max\{\delta_s,\delta_T\}>0.01$, **do**
3. $k=k+1$;
4. Solve problem (11), and obtain $\left(\mathbf{s}^{(k)},T^{\mathrm{ul},(k)}\right)$ via interior-point method;
5. Calculate $\mathbf{p}^{\mathrm{T},(k)}$ via Algorithm 1 with $\mathbf{p}^{\mathrm{T},(k-1)}$ and $\left(\mathbf{s}^{(k)},T^{\mathrm{ul},(k)}\right)$;
6. Calculate $\delta_s=\max\limits_n\left|\frac{s_n^{(k)}-s_n^{(k-1)}}{s_n^{(k-1)}}\right|$;
7. Calculate $\delta_T=\left|\frac{T^{\mathrm{ul},(k)}-T^{\mathrm{ul},(k-1)}}{T^{\mathrm{ul},(k-1)}}\right|$;
8. **end while**

TABLE I
SIMULATION PARAMETERS

Parameters	Typical value	Parameters	Typical value
Cell radius	500 m	M	100
N	32	Input bits	7×10^5
T	0.3s	λ	10^{-26}
ε	40	$f_{\max}$	800 MHz
$f_i^{\mathrm{BS}},i=1$	16 GHz	$f_i^{\mathrm{BS}},i>1$	8 GHz
$p_{\max}^{\mathrm{T}}$	23 dBm	Noise power	-174 dBm/Hz

is organized in **Algorithm 2**, whose *convergence* and *effectiveness* are further testified.

Proposition 2: Algorithm 2 monotonically decreases the objective function $\Upsilon\left(\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}\right)$ of problem (10) at each iteration, and eventually converges within finite iterations.

Proof: Considering the k-th iteration of Algorithm 2, and we have

$$\begin{aligned}&\Upsilon\left(\mathbf{s}^{(k-1)},\mathbf{p}^{\mathrm{T},(k-1)},T^{\mathrm{ul},(k-1)}\right)\\&\overset{(\vartheta)}{\geq}\Upsilon\left(\mathbf{s}^{(k)},\mathbf{p}^{\mathrm{T},(k-1)},T^{\mathrm{ul},(k)}\right)\\&\overset{(\varsigma)}{\geq}\Upsilon\left(\mathbf{s}^{(k)},\mathbf{p}^{\mathrm{T},(k)},T^{\mathrm{ul},(k)}\right),\end{aligned}\tag{19}$$

where the inequality (ϑ) holds because problem (11) is convex and $\left(\mathbf{s}^{(k)},T^{\mathrm{ul},(k)}\right)$ is its global optimal solution; the inequality (ς) is proved to be valid in Proposition 1. Therefore, $\Upsilon\left(\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}\right)$ is reduced at each iteration. Besides, as $\Upsilon\left(\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}\right)$ is lower-bounded, Algorithm 2 must converge in finite iterations for a given threshold. ■

As derived before, the power control solution can be found by solving a series of convex optimization problems with polynomial complexity. Besides, the subproblem for joint optimization of computational task offloading ratio and uplink transmission duration itself is a convex problem, which can be optimally solved with polynomial complexity required. Furthermore, as verified by Fig. 1, Algorithm 2 converges fast. In summary, the proposed alternating optimization algorithm only needs polynomial computational complexity.

IV. Simulation Results

In the simulation, a two-tier cell is considered with one MBS equipped with massive MIMO, three SBSs, and uniformly distributed users. The large-scale channel fading between BSs and users is composed of the pathloss (dB) $128.1+37.6\log_{10}d$ (km), and the shadow fading with standard derivation 8 dB. The rest parameters and their default values are given in Table I.

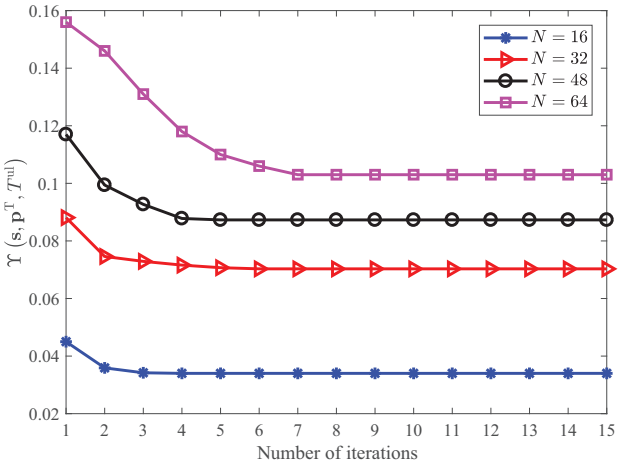

Fig. 1. The convergence procedure of Algorithm 2.

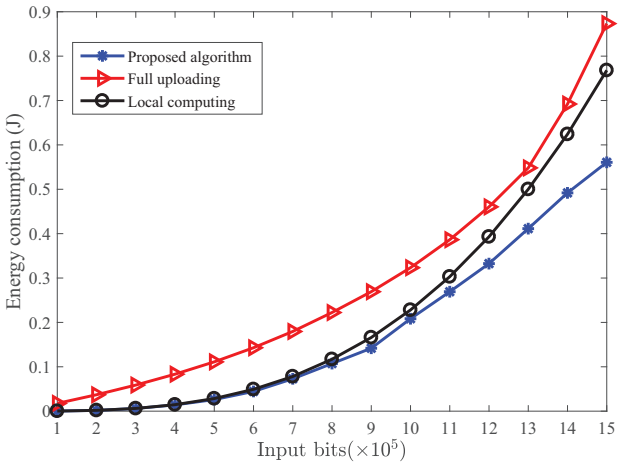

Fig. 2. Energy consumption *vs.* input data bits for different algorithms.

Since the proposed algorithm is a two-stage iterative algorithm, we first show the convergence performance of Algorithm 2 in Fig. 1. As shown in all the four curves, $\Upsilon\left(\mathbf{s},\mathbf{p}^{\mathrm{T}},T^{\mathrm{ul}}\right)$ decreases consistently and converges within only several iterations, which is in accordance with Proposition 1. Besides, it can be found that the number of users slightly influences the speed of convergence.

Then, the effectiveness of 'Proposed algorithm' is validated via simulation results shown in Figs. 2-3, in comparison with two benchmark schemes: 'Local computing' which means that all users execute their whole computation tasks locally; 'Full uploading' which indicates that the whole tasks of users are offloaded to their corresponding BSs and executed remotely.

It can be observed from Figs. 2-3 that all the three curves follow the same tendency, where the energy consumption

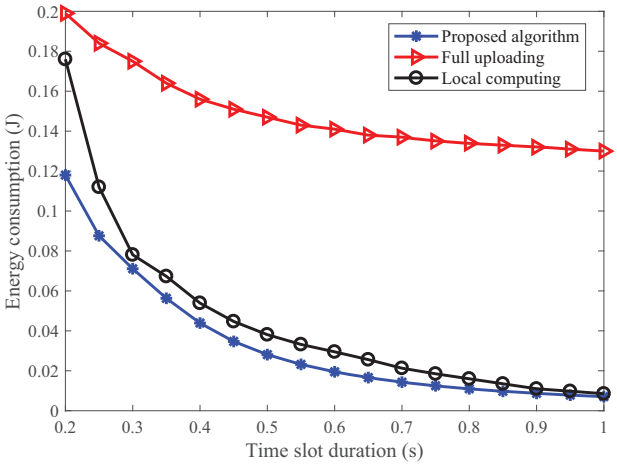

Fig. 3. Energy consumption *vs.* time slot duration for different algorithms.

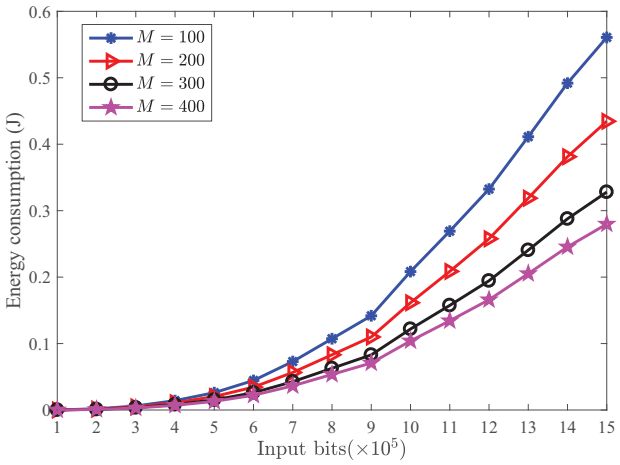

Fig. 4. Energy consumption *vs.* input data bits with different numbers of antennas.

of mobile devices increases with the input data bits but deceases with the time slot duration. It is worth noticing that our proposed algorithm significantly outperforms both 'Local computing' and 'Full uploading' schemes in two figures. When the number of input bits is small or the time slot is long enough, since 'Local computing' consumes much less energy than 'Full uploading', our proposed algorithm makes most of input bits executed locally, whose performance is therefore close to 'Local computing' as shown in Figs. 2-3. With the increase of input bits or the decrease of the time slot duration, the energy consumption of both basic schemes goes up rapidly. Under this circumstance, our proposed algorithm can flexibly adjust the ratio of uploading bits to avoid wasting energy on uplink transmission with bad channel or local computing with too high CPU frequency, which consequently saves more mobile energy.

Finally, Fig. 4 presents the impact of the antenna number M on the total energy consumption. We can find in Fig. 4 that larger number of active antennas corresponds to lower energy consumption, since increasing the antenna number equipped at the MBS helps to reduce the transmit power of users connected with the MBS. This indicates that the employment of massive MIMO helps lower the energy consumption of mobile devices.

V. Conclusion

In this paper, we have studied energy-efficient multi-user computation offloading problem in massive MIMO enabled HetNets, and proposed a low-complexity alternating optimization algorithm for the joint optimization of the CPU-cycle frequency of mobile devices, uplink power, computational task offloading ratio and uplink transmission duration. Through theoretical analysis, we have found that the proposed iterative algorithm converges within limited numbers of iterations. Numerical results verified the effectiveness of our algorithms as compared to local computing and full uploading schemes. Besides, it is shown that adopting massive MIMO in HetNets benefits mobile energy savings.

Acknowledgment

This work was supported in part by the Royal Society project IEC170324.

References

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

[2] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, Mar. 2017.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[4] Y. Hao, Q. Ni, H. Li, and S. Hou, "On the energy and spectral efficiency tradeoff in massive MIMO enabled HetNets with capactity-constrained backhaul links," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4720–4733, Nov. 2017.

[5] Z. Song, Q. Ni, K. Navaie, S. Hou, S. Wu, and X. Sun, "On the spectral-energy efficiency and rate fairness tradeoff in relay-aided cooperative OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6342–6355, Sep. 2016.

[6] Y. Hao, Q. Ni, H. Li, and S. Hou, "Robust multi-objective optimization for EE-SE tradeoff in D2D communications underlaying heterogeneous networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4936–4949, Oct. 2018.

[7] 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.

[8] 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.

[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] Y. Mao, J. Zhang, S. Song, and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[11] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.

[12] J. Zhang, W. Xia, F. Yan, and L. Shen, "Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing," *IEEE Access*, vol. 6, pp. 19 324–19 337, Mar. 2018.

[13] A. He, L. Wang, M. Elkashlan, Y. Chen, and K.-K. Wong, "Spectrum and energy efficiency in massive MIMO enabled HetNets: A stochastic geometry approach," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2294–2297, Dec. 2015.

[14] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, "Distributed energy efficient fair user association in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1770–1773, Oct. 2015.

[15] Y. Hao, Q. Ni, H. Li, and S. Hou, "Energy and spectral efficiency tradeoff with user association and power coordination in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2091–2094, Oct. 2016.

[16] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[17] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. USENIX Conf. Hot Topics Cloud Computi. (HotCloud)*, no. 1-7, Boston, MA, USA, Jun. 2010, pp. 1–6.

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

[19] A. Zappone, E. Bjornson, L. Sanguinetti, and E. Jorswieck, "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2844–2859, Jun. 2017.