# Energy-Efficient Task Offloading in Massive MIMO-Aided Multi-Pair Fog-Computing Networks

Kunlun Wang, *Member, IEEE*, Yong Zhou, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Long Shi, *Member, IEEE*, Wen Chen, *Senior Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*— The energy-efficient task offloading problem of a massive multiple-input multiple-output (MIMO)-aided fog computing system is solved, where multiple task nodes offload their computational tasks to be solved via a massive MIMO-aided fog access node to multiple processing nodes in the fog for execution. By considering realistic imperfect channel state information (CSI), we formulate a joint task offloading and power allocation problem for minimizing the total energy consumption, including both computation and communication power consumptions. We solve the resultant non-convex optimization problem in two steps. First, we solve the computational task allocation and computational resource allocation for a given power allocation. Then, we conceive a sequential optimization framework for determining the specific power allocation decision that minimizes the total energy consumption of the fog access node. Given the computational tasks, the computational resources, and the power allocation, we propose an iterative algorithm for the system optimization. The simulation results show that the proposed scheme significantly reduces the total energy consumption compared to the benchmark schemes.**

***Index Terms*— Fog computing, massive MIMO, computational task offloading, energy efficiency, fog access node.**

Manuscript received July 17, 2020; revised October 23, 2020 and December 13, 2020; accepted December 13, 2020. Date of publication December 21, 2020; date of current version April 16, 2021. The work of K. Wang was supported by the National Natural Science Foundation of China (NSFC) under grant 61801463. The work of Y. Zhou was supported by the NSFC under grants 62001294, U20A20159, and 61971286. This work of J. Li was supported by National Key R&D Program under Grants 2018YFB1004800, in part by NSFC under Grants 61727802,61872184. The work of W. Chen was supported by National Key Project 2018YFB1801102 and 2020YFB1807700, and by NSFC 61671294 and 62071296, and by STCSM 20JC1416502. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, EP/P003990/1 (COALESCE), of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom. This article has been submitted in part at the IEEE International Conference on Communications (ICC), Montreal, Canada, June 14–18, 2021. The associate editor coordinating the review of this article and approving it for publication was T. He. *(Corresponding author: Yong Zhou.)*

Kunlun Wang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and also with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China (e-mail: wangkl2@shanghaitech.edu.cn).

Yong Zhou is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: zhouyong@shanghaitech.edu.cn).

Jun Li and Long Shi are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jun.li@njust.edu.cn; slong1007@gmail.com).

Wen Chen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenchen@sjtu.edu.cn).

Lajos Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2020.3046265.

Digital Object Identifier 10.1109/TCOMM.2020.3046265

## I. Introduction

GIVEN the rapid development of the Internet of Things (IoT), more and more intelligent things and smart objects are being connected to the network [1], [2]. Meanwhile, the improved networking speed paves the way for sophisticated multitasking applications, such as online gaming, augmented reality and space-air-ground services [3], [4]. These novel applications typically require low latency and demand prompt computations energy mangement for realtime task processing and high data rates. However, given their compact form-factor, mobile devices have limited computation, storage, and energy resources. To overcome these limitations, fog computing, as an emerging technology, has been proposed for sticking a compelling compromise between the resource-constrained nature of compact devices and their high-complexity tasks. As a result, fog computing is capable of significantly reducing the computing burden of the mobile devices by efficiently utilizing the abundant computation resources in the fog around them, which is taking advantage of pervasive mobile devices and their pairwise encounters to form a pool of computation resources.

However, in fog computing, a user offloads his/her computing task to a server in the uplink, and the processed data has to be sent back to the user in the downlink. Hence, the performance of fog computing operation also depends on the communication performance. With the advent of the fifth generation (5G) wireless standards, new high performance technologies have been introduced. One of these key technologies is constituted of massive Multiple-input multiple-output (MIMO) systems [5], [6], which are being increasingly adopted in different networking and computing frameworks. The authors in [7]–[10] mainly consider single-antenna systems taking joint wireless resources and task offloading into account and fail to exploit the advantages brought by MIMO technology in terms of offloading efficiency. MIMO techniques have the potential of achieving high channel capacity [11]–[13]. New technologies are being introduced to improve the performance of mobile users from current levels. By equipping the base stations (BSs) with a

large number of antennas, widely known as massive MIMO, has emerged as one of the most promising solutions [14], [15] that significantly enhance the systems spectral efficiency (SE) and energy efficiency (EE) trade-off. More specifically, when the number of antennas increases, the channels become more deterministic, which is referred to as channel hardening. Hence, the achievable data rates are mostly determined by large-scale fading, and so is the resource allocation. This means that there is no need to frequently update the resource allocation, yielding substantial savings in the signalling overhead. In all, massive MIMO schemes increase the spectral and energy efficiencies and support an increased number of users, both of which are crucial for fog computing systems.

Additionally, relay techniques have been integrated into various wireless communication standards to improve the coverage and throughput [16]. MIMO relay networks facilitate spectral-efficient, and reliable data transmission between spatially distributed user nodes and multi-antenna destinations via intermediate multi-antenna relay nodes [17], [18]. In this work, we propose to use a massive MIMO-aided fog access node (FAN) regarded as a relay for significantly improving the data rate of computational task offloading as well as the task execution efficiency.

### A. Related Works

In recent years, task offloading has gained increasing attention in a diverse range of fog computing scenarios [8], [9], [19]–[23]. In contrast to the traditional cloud-based computing architecture, fog computing provides a more efficient platform for low-latency task offloading at a high energy-efficiency. In particular, Li *et al.* [24] explored the concept of mobile cloudlets in mobile cloud computing through studying the cloudlet properties and the computing performance attained. Xiao *et al.* [25] proposed a task offloading framework for a mobile user, who may offload computing tasks to the nearby devices for exploiting the processing capacity available in the vicinity. Chen and Zhang [26] proposed a hybrid task offloading framework to support both device-to-device (D2D) offloaded execution and cloud offloaded execution. As a further development, Pu *et al.* [7] formulated an optimization problem for minimizing the time-averaged energy consumption for task execution of all users in D2D networks. Wang *et al.* [20] designed an online learning based task offloading algorithm for delay-sensitive applications in dynamic fog networks by exploiting the combinatorial multi-armed bandits (CMAB) framework. Chen *et al.* [27] developed a novel framework that enables the implementation of federated learning algorithms over wireless network, which jointly considers power- and computational-resource allocation. Yang *et al.* [9] proposed an energy-efficient fog computing framework associated with multiple neighboring helper nodes sharing their computational resources, taking into account the opportunistic spectrum access for spectrum sharing. Yang *et al.* [28] investigated a fundamental multi-task multi-helper problem in heterogeneous fog networks, i.e., how to effectively associate task nodes and helper nodes to minimize the delay of every task. However, since most of the contributions on task offloading have been focused on the single-hop resource allocation, there is a paucity of literature on multi-hop scenarios.

Resource allocation for peer offloading in fog-assisted small cell networks has been widely studied. Zhou *et al.* [29] proposed an online distributed task offloading (DTO) algorithm for practical fog computing systems, where each mobile user dynamically offloads its decision to nearby mobile devices in a collaborative manner via peer-to-peer wireless communications. Chen *et al.* [30] investigated peer offloading schemes in mobile edge computing-aided small cell networks, where diverse task arrival patterns are considered both in the spatial and temporal domains. Although the above outstanding contributions have studied peer-to-peer computation offloading in single-antenna systems, the potential benefits of massive MIMO schemes in further enhancing the performance of the fog computing framework have not been explored. Different from the conventional MIMO, massive MIMO configuration relies on a large number of antennas and can significantly improve the data rate of task offloading. It has been shown in [31] that massive MIMO schemes significantly improve the data rates of user equipments (UEs) at the cell edge, as well as the overall network throughput. As expected, the integration of fog computing and massive MIMO can enhance the performance of task offloading in multi-user fog computing systems [32]–[35]. In particular, Bursalioglu *et al.* [32] proposed and analyzed an architecture nicknamed fog massive MIMO, where a large number of multi-antenna BSs are densely deployed, and serve the users using zero-forcing beamforming (ZFBF). In [33], Pirzadeh *et al.* investigated the viability of supervised-learning methods in estimating the user locations by observing across the fog massive MIMO network signals transmitted by the users. In [34], Chen proposed a specific fog computing mechanism for the uplink of fronthaul-constrained distributed massive MIMO systems (DM-MIMO), and the corresponding power control algorithm. Mungara *et al.* [35] considered a new architecture underpinned by, on-the-fly, pilot contamination control, termed as fog massive MIMO, where the users are able to establish high-throughput and low-latency data links in a seamless and opportunistic manner, as they travel through a dense fog of high capacity remote radio heads (RRHs). Although the aforementioned studies have demonstrated the benefits of massive MIMO-based fog computing, they have not taken into account the channel estimation error in resource allocation and task offloading, which are particularly important for time-variant fog computing systems. On the other hand, as the fog systems provide additional computing capabilities at the edge of the network, a major question that they raise is how to manage task execution. More precisely, how to decide which tasks to be executed in the end-users stratum, the fog stratum, and the cloud stratum. On a finer level, the dilemma is which nodes a particular task should be assigned to.

### B. Main Contributions

In this contribution, we jointly optimize the task offloading and power allocation of the massive MIMO-aided fog computing systems, where robust resource allocation is conceived in

the face of realistic channel estimation errors. In our proposed fog computing framework, some of the nodes referred to the parlance as task nodes (TNs) have computationally-intensive applications to run, which hence request the offloading of their computational tasks via a massive MIMO-enabled FAN to computing nodes (CNs) having under-utilized computational resources. Again, we assume that the downlink (DL) channel state information (CSI) of data transmission from the FAN to the CN (FAN-CN) is imperfect. Then, we extend to the imperfect CSI assumption to the link spanning from the TNs to the FAN (TN-FAN), namely to the uplink (UL). After establishing the total task offloading energy consumption, we formulate a joint task offloading and power allocation problem. The objective is to minimize the total energy consumption, while taking into account the practical communication and computation constraints. Since the optimization problem is non-convex, it is challenging to obtain an optimal solution. Additionally, considering imperfect CSI further complicates the optimization problem. To this end, we solve the task offloading and power allocation problem in two steps. First, we determine the task and computation resource allocation for given power allocation results. Then, we present a sequential optimization framework for determining the power allocation decision that minimizes the total energy consumption at the TNs and the FAN. Based on the task-, computational resource-, and power-allocations, we propose an iterative algorithm for finding the jointly optimized results. The main contributions of this paper are summarized as follows.

- We develop a novel massive MIMO-enabled task offloading framework, where multiple nodes offload their computational tasks to multiple CNs via a massive MIMO-aided FAN. We formulate an energy minimization problem by jointly optimizing the allocation of tasks, computational resource, and power.
- We partition the original optimization problem into two subproblems, namely into, task and computational resource allocation subproblem and a FAN power allocation subproblem. In the optimization problem, we first consider realistic imperfect FAN-CN CSI and then we extend to the imperfect TN-FAN CSI to obtain the robust power allocation results.
- We formulate the power allocation subproblem as a non-convex problem when the computational resource allocations and power allocations of each node having computational tasks are fixed, and present a sequential optimization framework for carrying out the power allocation decisions. Based on the task, computational resource, and power allocations, we propose an iterative algorithm for finding the jointly optimized results. Furthermore, we prove the convergence of the proposed iterative algorithm.
- Our simulation results demonstrate that the proposed computational task offloading and power allocation algorithm achieves significant performance improvements over the benchmarks.

The rest of the paper is organized as follows. Section II describes our system model and problem formulation. Then, we formulate a task offloading and power allocation problem. In Section III, we introduce the total energy consumption of our massive MIMO fog computing systems. In Section IV, we optimize the task offloading, computational resource allocation, and power allocation by proposing an iterative optimization algorithm for massive MIMO-aided fog computing networks. In Section V, we discuss our simulation results. Finally, our conclusions are provided in Section VI. Table I lists the frequently used notations.

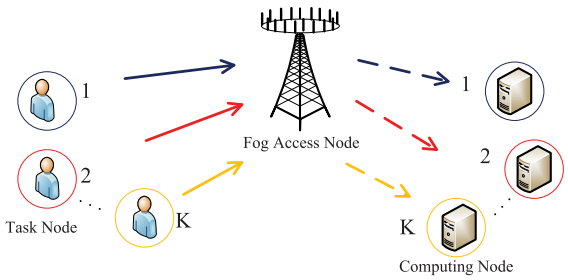



Fig. 1. Illustration of a massive MIMO-aided fog computing network, where $K$ task nodes offload their tasks to $K$ computing nodes in the fog with the aid of a fog access node relying on a massive MIMO scheme.

Matrices and vectors are denoted by capital and lower-case boldface letters, respectively. $\mathbb{C}^{M\times N}$ and $\mathbb{R}^{M\times N}$ denote the sets of all $M \times N$ complex-valued matrix and real-valued matrix, respectively. $(\cdot)^{\mathrm{H}}$, $(\cdot)^{\dagger}$, $\mathrm{tr}(\cdot)$ and $\mathsf{E}(\cdot)$ denote the conjugate transpose, pseudo-inverse, trace and the expectation, respectively. i.i.d. stands for independent and identically distributed.

## II. System Model and Problem Formulation

In this section, we first describe the network model of the massive MIMO-enabled fog computing networks, present the channel model as well as computational model, and then formulate the total energy minimization problem.

### A. Network Model

We consider a massive MIMO aided fog computing network, consisting of $K$ single-antenna TNs, an $M$-antenna FAN, and $K$ single-antenna CNs, as shown in Fig. 1. Each TN can either offload its task to the intended CN via the FAN or execute the computational task locally. The multi-antenna FAN serves as a relay to help offloading the tasks from the TNs to the CNs. Due to the associated signal decoding and resource scheduling complexities of non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA), the extension to consider task offloading from multiple task nodes to multiple computing nodes is left for future discussions.

The fog computing system operates over a bandwidth of $B$ and the time is slotted into intervals of constant duration $T$. We assume that each CN can only execute one task from a single TN during each time slot. Without loss of generality, we also assume that the $k$th TN is paired with the $k$th CN for task offloading, which is hence referred to as the $k$th TN-CN pair. To reduce the offloading delay, the task offloading

TABLE I
FREQUENTLY USED NOTATIONS

| Definition | Notation | Definition | Notation |
|---|---|---|---|
| Number of TNs | $K$ | Number of antennas of FAP | $M$ |
| Channel coefficient matrix from the $K$ TNs to the FAN | $\mathbf{H}$ | CPU cycle frequency of TN $k$ | $f_k$ |
| Channel coefficient matrix from the FAN to the $K$ CNs | $\mathbf{G}$ | Amount of task | $b_k$ |
| Ratio of data bits offloaded to the total task bits | $\nu_k$ | Number of cycles needed for computing each single data bit | $\epsilon$ |
| Energy consumption of local computing at TN $k$ | $E_k^L$ | Power consumption of TN $k$ | $P_k^L$ |
| Time duration of the task offloading from TN $k$ | $D_k$ | Average task offloading rate for the $k$th TN | $\mathcal{R}_k$ |
| Energy consumption for the task offloading of TN $k$ | $E_k^{\text{off}}$ | FAN transmit power allocated to the $k$th TN-CN pair | $p_k$ |
| Total energy consumption of task offloading of TN $k$ | $E_{\text{total,k}}$ | Transmit power of each TN | $P_t$ |
| SINR constraint | $\gamma_0$ | maximum CPU-clock frequency of TN $k$ | $f_{\max}$ |
| Transmit symbols of all the TNs | $\mathbf{x}$ | Symbol delivered from the $k$th TN to its paired CN | $s_k$ |
| Signal received at the FAN | $\mathbf{y}_R$ | Additive white Gaussian noise at the FAN | $\mathbf{n}_R$ |
| Precoding matrix of the FAN | $\hat{\mathbf{W}}$ | Signal received at all CNs | $\mathbf{y}_U$ |

procedures of all TN-CN pairs are performed simultaneously. Additionally, there is no direct link between any TN and CN due to propagation obstacles. The FAN relies on time-division duplexing. In this context, the task offloading from the TNs to CNs consists of three phases, namely, the channel estimation phase, the task uploading phase from the TNs to the FAN (i.e., TN $\rightarrow$ FAN phase), and the task relaying phase from the FAN to the CNs (i.e., FAN $\rightarrow$ CN phase).

### B. Channel Model

We consider independent and identically distributed (i.i.d.) quasi-static Rayleigh fading. In particular, each inter-node channel remains invariant within one time slot, but varies independently across different time slots and links. As shown in [36], the assumption of i.i.d. Rayleigh fading permits the derivation of exact (non-asymptotic) ergodic capacity lower bounds for very comprehensive Massive MIMO systems, and experiments have established conditions under which this model is approximately valid [37]. Let $\mathbf{H} = [\mathbf{h}_1^{\mathrm{T}}, \cdots, \mathbf{h}_K^{\mathrm{T}}] \in \mathbb{C}^{M\times K}$ denote the $M \times K$ channel coefficient matrix from the $K$ TNs to the FAN, where the $k$th element $\mathbf{h}_k$ denotes the channel coefficient vector between the $k$th TN and FAN, $k = 1, 2, \ldots, K$. Additionally, let $\mathbf{G} = [\mathbf{g}_1^T, \cdots, \mathbf{g}_K^T] \in \mathbb{C}^{K\times M}$ denote the $K \times M$ channel coefficient matrix from the FAN to the $K$ CNs, where the $k$th element $\mathbf{h}_{D,k}$ denotes the channel coefficient vector between the FAN and the $k$th CN, $k = 1, 2, \ldots, K$.

It is reasonable to assume that the CSI is perfectly known at the receiver [38], since the receiver can acquire the accurate CSI at the receiver (CSIR) with training. However, the transmitter can only acquire the imperfect CSI through a finite-rate feedback channel, which introduces quantization error and feedback delay. Consequently, we assume that the CSI in the TN$\rightarrow$FAN phase is perfectly known at the FAN as the UL receiver, while the CSI in the FAN$\rightarrow$CN phase is imperfectly known at the FAN as the DL transmitter. Let $\hat{\mathbf{G}}$ denote the estimated FAN-CN channel CSI. In this context, the FAN-CN channel can be modeled as [39]

$$\mathbf{G} = \sqrt{1-\tau_D^2}\hat{\mathbf{G}} + \tau_D \mathbf{\Omega}_D, \tag{1}$$

where $\mathbf{\Omega}_D \in \mathbb{C}^{K\times M}$ has i.i.d entries with zero mean and unit variance independent of the estimated channel matrix $\hat{\mathbf{G}}$, and the parameter $\tau_D \in [0, 1]$ reflects the estimation accuracy or quality of the channel matrix $\mathbf{G}$. The case of $\tau_D = 0$ corresponds to perfect CSI estimation, whereas the CSI is completely unknown if $\tau_D = 1$.

### C. Computation Models

In this subsection, we discuss both the local and the CN computing approaches.

*1) Partial Computing Offloading:* Let us consider that TN $k$ has $b_k$ bits to be computed in a time slot. Let us furthermore denote the ratio of data bits offloaded to the total task bits by $\nu_k$, i.e., $(1-\nu_k)b_k$ bits are subject to local computing and $\nu_k b_k$ to CN computing.

For local computing, the power consumption of TN $k$ can be modeled as [40]

$$P_k^L = \varrho f_k^3, \tag{2}$$

where $f_k$ is the CPU cycle frequency of TN $k$, which can be adjusted via the dynamic voltage and frequency scaling (DVFS) technique [41]. Thus, the local computing time of TN $k$ is calculated as

$$t_k^L = \frac{\epsilon(1-\nu_k)b_k}{f_k}, \tag{3}$$

where $\epsilon$ ($\epsilon > 0$) denotes the number of cycles needed for computing each single data bit.[1] Consequently, the energy consumption of local computing at TN $k$ is given by

$$E_k^L = P_k^L t_k^L = \varrho\epsilon(1-\nu_k)b_k f_k^2. \tag{4}$$

As for computing by the CN, the TNs first offload their tasks to the FAN. After collecting the input bits from the TNs, the FAN distributes the tasks to the corresponding CNs. Let us define the time duration of the task offloading from the TN $k$ to its intended CN by $D_k$. Thus, the energy consumption of TN $k$ for task offloading is expressed as

$$E_k^T = P_t D_k, \tag{5}$$

where $P_t$ is the transmit power of each TN. We assume furthermore that the time and energy consumptions of the TNs required to download the computed results are negligible, since the computing results are usually of small size and the FAN has a high transmit power [42]–[44].

[1]Note that the parameters $b_k$ and $\epsilon$ are determined by the types of applications and estimated via task profilers [42].

*2) Task Offloading:* Task offloading refers to the case that the task is offloaded for execution by the paired CN. The task offloading time from TN $k$ to its intended CN is given by

$$D_k = \frac{\nu_k b_k}{\mathcal{R}_k}, \quad \forall k, \tag{6}$$

where $\mathcal{R}_k$ is the average task offloading rate for the $k$th TN. Correspondingly, the energy consumption for the task offloading of TN $k$ is given by

$$E_k^{\text{off}} = (P_t + p_k)D_k = \frac{(P_t + p_k)\nu_k b_k}{\mathcal{R}_k}, \tag{7}$$

where $p_k \in \mathbf{p} = [p_1, \cdots, p_K]$ is the FAN transmit power allocated to the $k$th TN-CN pair. Let $P_r$ denote the maximum transmit power available at the FAN. As such, we have $p_k \leq P_r$.

Upon receiving the computational tasks, the CN allocates its computational resources for task execution. In this context, the total energy consumption of task offloading consists of the local computing energy consumption and the task offloading energy consumption. After combining (4)-(7), the total energy consumption of task offloading of TN $k$ is given by

$$E_{\text{total,k}} = \varrho\epsilon(1-\nu_k)b_k f_k^2 + \frac{(P_t + p_k)\nu_k b_k}{\mathcal{R}_k}. \tag{8}$$

### D. Problem Formulation

In this section, we formulate a joint task-, power-, and computational-resource allocation problem with an objective of minimizing the total energy consumption, taking into account both the communication and computational constraints. Let $P_{t_{\max}}$ in (9d) and $\gamma_k$ in (9e) denote the maximum transmit power of each TN and the received signal-to-interference-plus-noise ratio (SINR) at CN $k$, respectively. To minimize the total energy consumption $E_{\text{total}}$ of $K$ TNs and FAN while ensuring that their tasks are successfully executed within a single time slot, the energy-efficient multi-pair computation offloading problem is formulated as

$$\min_{\boldsymbol{\nu},\mathbf{p},\mathbf{f}} \; E_{\text{total}} \tag{9a}$$
$$\text{s.t. } 0 \leq \nu_{\text{k}} \leq 1, \quad \forall k, \tag{9b}$$
$$0 \leq p_k \leq P_r, \quad \forall k, \tag{9c}$$
$$0 \leq P_t \leq P_{t_{\max}}, \tag{9d}$$
$$\gamma_k \geq \gamma_0, \quad \forall k, \tag{9e}$$
$$0 \leq f_k \leq f_{\max}, \quad \forall k, \tag{9f}$$
$$\frac{\epsilon(1-\nu_k)b_k}{f_k} \leq T, \quad \forall k, \tag{9g}$$

where (9b) gives the range of the computational task offloading ratio; (9c) specifies the power allocation variables for the FAN; (9d) is the transmit power constraint for the TN; (9e) is the quality-of-service (QoS) of delay constraints capable of ensuring that the SINR of each TN-CN pair is higher than $\gamma_0$; (9f) represents that the maximum CPU-clock frequency of TN $k$ is $f_{\max}$.

## III. Energy Consumption Analysis

This section investigates the total energy consumption of the massive MIMO-aided fog computing systems. Firstly, we derive the received SINR for determining the offloading rate. Secondly, the task offloading time is calculated. Finally, the total energy consumption is analyzed.

### A. Task Transmission

As for the task computation, a CN can execute either all tasks after receiving all of them or some tasks while still receiving more tasks. Given the overlapped arrival order of tasks at the CN, the overlapping nature of the computing task makes the analysis intractable. For simplicity, let us assume that each CN only executes the task received from the intended TN after receiving all tasks. As a result, we consider that task transmission in massive MIMO-aided fog computing networks consists of TN $\rightarrow$ FAN phase and FAN $\rightarrow$ CN phase.

In the TN $\rightarrow$ FAN phase, all TNs simultaneously transmit their symbols to the FAN in a single time slot, which is given by

$$\mathbf{x} = \sqrt{P_t}\mathbf{s}, \tag{10}$$

where $\mathbf{s} = [s_1, \cdots, s_K]^T$ is an information-bearing symbol vector with $\mathsf{E}(\mathbf{s}\mathbf{s}^\dagger) = \mathbf{I}_K$, and $s_k$ is the symbol delivered from the $k$th TN to its paired CN. The signal $\mathbf{y}_R \in \mathbb{C}^{M\times 1}$ received at the FAN is

$$\mathbf{y}_R = \mathbf{H}\mathbf{x} + \mathbf{n}_R, \tag{11}$$

where $\mathbf{n}_R \in \mathbb{C}^{M\times 1}$ is the zero-mean additive white Gaussian noise (AWGN) at the FAN with a variance of $\mathsf{E}(\mathbf{n}_R\mathbf{n}_R^{\text{H}}) = \sigma_r^2\mathbf{I}_M$. Given the knowledge of perfect CSIR and imperfect CSIT, the FAN precodes its received signal $\mathbf{y}_R$ and obtains the filtered signal vector $\mathbf{x}_R \in \mathbb{C}^{M\times 1}$ as

$$\mathbf{x}_R = \hat{\mathbf{W}}\mathbf{y}_R, \tag{12}$$

where $\hat{\mathbf{W}} \in \mathbb{C}^{M\times M}$ is the precoding matrix. The precoding matrix of the FAN can be written as

$$\hat{\mathbf{W}} = \hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger, \tag{13}$$

where $\hat{\mathbf{G}}^\dagger = \hat{\mathbf{G}}^{\text{H}}(\hat{\mathbf{G}}\hat{\mathbf{G}}^{\text{H}})^{-1}$ and $\mathbf{H}^\dagger = (\mathbf{H}^{\text{H}}\mathbf{H})^{-1}\mathbf{H}^{\text{H}}$. The diagonal matrix $\mathbf{P} \in \mathbb{R}^{K\times K}$ is the power allocation matrix of the FAN, wherein the $k$th diagonal element $[\mathbf{P}]_{k,k} = \sqrt{p_k}$ denotes the transmit power allocated to the $k$th TN-CN pair. The average power constraint at the FAN can be written as

$$\mathsf{E}[\text{tr}(\mathbf{x}_{\text{R}}\mathbf{x}_{\text{R}}^{\text{H}})] \leq P_r. \tag{14}$$

### B. Received SINR at CN

During the FAN $\rightarrow$ CN phase, the FAN broadcasts $\mathbf{x}_R$ to all the $K$ active CNs. The signal received at all CNs is given by

$$\mathbf{y}_U = \mathbf{G}\hat{\mathbf{W}}\mathbf{y}_R + \mathbf{n}_U, \tag{15}$$

where $\mathbf{y}_U = [y_1, \cdots, y_K] \in \mathbb{C}^{K\times 1}$, and $\mathbf{n}_U$ is the zero-mean AWGN at the destinations with a variance of $\mathsf{E}(\mathbf{n}_U\mathbf{n}_U^{\text{H}}) = \sigma_U^2\mathbf{I}_K$.

Given (15), the signal vector received at all the CNs can be rewritten as

$$\begin{aligned}\mathbf{y}_U &= \mathbf{G}\hat{\mathbf{W}}\mathbf{H}\mathbf{x} + \mathbf{G}\hat{\mathbf{W}}\mathbf{n}_R + \mathbf{n}_U \\ &= \mathbf{G}\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger\mathbf{H}\mathbf{x} + \mathbf{G}\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger\mathbf{n}_R + \mathbf{n}_U \\ &= (\sqrt{1-\tau_D^2}\hat{\mathbf{G}} + \tau_D\mathbf{\Omega}_D)\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{x} \\ &\quad + (\sqrt{1-\tau_D^2}\hat{\mathbf{G}} + \tau_D\mathbf{\Omega}_D)\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger\mathbf{n}_R + \mathbf{n}_U \\ &= \sqrt{1-\tau_D^2}\mathbf{P}\mathbf{x} + \tau_D\mathbf{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{x} + \sqrt{1-\tau_D^2}\mathbf{P}\mathbf{H}^\dagger\mathbf{n}_R \\ &\quad + \tau_D\mathbf{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger\mathbf{n}_R + \mathbf{n}_U. \end{aligned} \tag{16}$$

Based on (16), the signal received at the $k$th CN is

$$\begin{aligned}y_{U,k} = \sqrt{p_k P_t}s_k + \sqrt{p_k}\hat{\mathbf{f}}_{S,k}^\dagger\mathbf{n}_R + \sqrt{p_k P_t}\hat{\mathbf{f}}_{S,k}^\dagger\tilde{\mathbf{F}}_S s \\ + \tilde{\mathbf{f}}_{D,k}\mathbf{x}_R + n_k, \end{aligned} \tag{17}$$

where $y_{U,k}$, $s_k$, and $n_k$ are the $k$th elements of $\mathbf{y}_U$, $\mathbf{s}$, and $\mathbf{n}_U$, respectively, and $\hat{\mathbf{f}}_{S,k}^\dagger$ and $\tilde{\mathbf{f}}_{D,k}$ are the $k$th row of $\hat{\mathbf{F}}_S^\dagger$ and $\tilde{\mathbf{F}}_D$ respectively. The effective SINR of the $k$th data stream at the CN is given by

$$\gamma_k = \frac{(1-\tau_D^2)p_k P_t}{P_t(\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^H) + \sigma_r^2(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^H) + \sigma_u^2}, \tag{18}$$

where $\boldsymbol{\zeta}_k$ and $\boldsymbol{\eta}_k$ are the $k$th rows of $\tau_D\mathbf{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}$ and $(\sqrt{1-\tau_D^2}\mathbf{P}\mathbf{H}^\dagger + \tau_D\mathbf{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\mathbf{H}^\dagger)$, respectively. In the following theorem, we characterize the asymptotic property of the SINR in (18) under the massive MIMO setting.

*Theorem 1: As the number of antennas at the FAN tends to $M \to \infty$, the effective SINR in* (18) *can be asymptotically expressed as*

$$\gamma_{k,\infty} = \frac{(1-\tau_D^2)p_k P_t}{(1-\tau_D^2)\sigma_r^2\lambda_k\sum_{i=1}^K p_i + \sigma_u^2}. \tag{19}$$

*Proof:* Please refer to Appendix A. □

### C. Offloading Time and Energy Consumption

Given (18), the task offloading rate of the $k$th TN is given by

$$\mathcal{R}_k = \frac{B}{2}\log_2(1+\gamma_k), \tag{20}$$

where $B/2$ indicates that the FAN works in the half-duplex mode. Then, based on (6) and (20), the offloading time of the $k$th TN is given by

$$D_k = \frac{2\nu_k b_k}{B\log_2(1+\gamma_k)}. \tag{21}$$

The total transmit energy consumption is given by that of the TNs and the FAN.[2] According to the transmit power consumption of the TN and FAN, the corresponding offloading energy consumption is given by

$$E_k^{\text{off}} = (P_t + p_k)D_k = \frac{2(P_t+p_k)\nu_k b_k}{B\log_2(1+\gamma_k)}. \tag{22}$$

[2]Following a practical power consumption model in [9], we assume that the circuit power is a constant, accounting for the fixed power consumption for controlling, site-cooling, and the load-independent power of baseband processors. To simplify the problem, we only consider the transmit power consumption.

Given the energy consumptions of the local computing and task offloading in (4) and (22), we obtain the total energy consumption of the massive MIMO-enabled fog computing system as

$$\begin{aligned}E_{\text{total}} &= \sum_{k=1}^K \left(E_k^L + E_k^{\text{off}}\right) \\ &= \sum_{k=1}^K \left(\varrho\epsilon(1-\nu_k)b_k f_k^2 + \frac{2(P_t+p_k)\nu_k b_k}{B\log_2(1+\gamma_k)}\right). \end{aligned} \tag{23}$$

## IV. Joint Radio and Computational Resource Optimization

In this section, we jointly optimize the task offloading, computational resource and transmit power allocations for minimizing the total energy consumption at the TNs and FAN. Firstly, we solve the subproblem of task- and computational-resource allocation. Secondly, we solve the subproblem of FAN power allocation. Finally, we optimize the joint problem by conceiving an iterative algorithm.

### A. Task- and Computational-Resource Allocation

In this subsection, we solve the task-scheduling subproblem to obtain the task- and computational-resource allocation under a fixed FAN power allocation. In the following, we transform the non-convex optimization problem of (9) into a tractable convex one.

Firstly, it can be verified that the objective function (OF) of Problem (9) monotonically increases with $f_k$, $\forall k$. Secondly, based on constraint (9g), we have $f_k \geq \frac{\epsilon(1-\nu_k)b_k}{T}$. Finally, the optimal CPU-cycle frequency of TN $k$ can be obtained as

$$f_k^\star = \frac{\epsilon(1-\nu_k)b_k}{T}. \tag{24}$$

By substituting (24) into (9), Problem (9) is equivalently transformed into

$$\min_{\boldsymbol{\nu},\mathbf{p}} \sum_{k=1}^K \frac{\varrho\epsilon^3(1-\nu_k)^3 b_k^3}{T^2} + \frac{2(P_t+p_k)\nu_k b_k}{B\log_2(1+\gamma_k)} \tag{25a}$$

$$\text{s.t. } 0 \leq \nu_{\text{k}} \leq 1, \quad \forall k, \tag{25b}$$

$$0 \leq p_k \leq P_r, \quad \forall k, \tag{25c}$$

$$0 \leq P_t \leq P_{t_{\max}}, \tag{25d}$$

$$\gamma_k \geq \gamma_0, \quad \forall k. \tag{25e}$$

Nevertheless, the transformed problem (25) is still non-convex. Next, we further divide it into two sub-problems of task- and computational-resource allocation and FAN power allocation, and solve them alternately.

The subproblem of task allocation with respect to the computational task offloading ratio is given by

$$\min_{\boldsymbol{\nu}} \ \varphi(\boldsymbol{\nu}) = \sum_{k=1}^K \frac{\varrho\epsilon^3(1-\nu_k)^3 b_k^3}{T^2} + \frac{2(P_t+p_k)\nu_k b_k}{B\log_2(1+\gamma_k)} \tag{26a}$$

$$\text{s.t. } 0 \leq \nu_{\text{k}} \leq 1, \quad \forall k. \tag{26b}$$

Problem (26) is convex and can be solved by using standard algorithms, such as the classic interior-point method at a polynomial computational complexity [45]. By taking the derivative of the objective function with respect to $s_k$, we have

$$\frac{\partial\varphi(\boldsymbol{\nu})}{\partial\nu_k}=\frac{-3\varrho\epsilon^3(1-\nu_k)^2b_k^3}{T^2}+\frac{2(P_t+p_k)b_k}{B\log_2(1+\gamma_k)}=0. \quad (27)$$

According to (27), we arrive at the optimal solution $\nu_k^*=1-\sqrt{\frac{2(P_t+p_k)b_kT^2}{3\varrho\epsilon^3b_k^3B\log_2(1+\gamma_k)}}$.

### *B. Power Allocation based on Sequential Optimization*

In this subsection, we propose a sequential optimization method for the FAN power allocation. By fixing the computational task offloading ratio vector $\boldsymbol{\nu}$, we only have to solve the power allocation problem. Thus, problem (25) can be simplified to

$$\min_{\mathbf{p}}\ \sum_{k=1}^{K}\frac{(P_t+p_k)\nu_kb_k}{\mathcal{R}_k}=\sum_{k=1}^{K}\frac{(P_t+p_k)\nu_kb_k}{\frac{B}{2}\log_2(1+\gamma_k)} \quad (28a)$$

$$\text{s.t. (9c), (9e).} \quad (28b)$$

Due to the non-convex OF and constraints, Problem (28) is still intractable. Next we use Theorem 1 to make the problem solvable. According to (19), we have

$$\begin{aligned}\gamma_k\le\gamma_{k,\infty}&=\frac{(1-\tau_D^2)p_kP_t}{(1-\tau_D^2)\sigma_r^2\lambda_k\sum_{i=1}^{K}p_i+\sigma_u^2}\\&\le\frac{(1-\tau_D^2)p_kP_t}{(1-\tau_D^2)\sigma_r^2\lambda_kp_k+\sigma_u^2}\\&=\frac{P_t}{\sigma_r^2\lambda_k}-\frac{\frac{P_t\sigma_u^2}{(1-\tau_D^2)\sigma_r^4\lambda_k^2}}{p_k+\frac{\sigma_u^2}{(1-\tau_D^2)\sigma_r^2\lambda_k}}.\end{aligned} \quad (29)$$

Now $\gamma_k$ becomes a concave function of $p_k$. To begin with the problem optimization, the OF of Problem (28) can be rewritten in form of a single ratio as

$$\sum_{k=1}^{K}\frac{(P_t+p_k)\nu_kb_k}{\mathcal{R}_k}=\sum_{k=1}^{K}\frac{(P_t+p_k)\nu_kb_k}{\frac{B}{2}\log_2(1+\gamma_k)}=\frac{\phi(\mathbf{p})}{\varphi(\mathbf{p})}, \quad (30)$$

where $\phi(\mathbf{p})=\sum_{k=1}^{K}\left[(P_t+p_k)\nu_kb_k\right]\prod_{l\ne k}\left[\frac{B}{2}\log_2(1+\gamma_l)\right]$ and $\varphi(\mathbf{p})=\prod_{k=1}^{K}\left[\frac{B}{2}\log_2(1+\gamma_k)\right]$, respectively.

Then, the Sequential Parametric Convex Approximation (SPCA) method of [46] can be applied to solve Problem (28), which can be transformed into the following problem

$$\min_{\mathbf{p}}\ \frac{\phi(\mathbf{p})}{\varphi(\mathbf{p})} \quad (31a)$$

$$\text{s.t. (9c),} \quad (31b)$$

$$\gamma_0-\left(\frac{P_t}{\sigma_r^2\lambda_k}-\frac{\frac{P_t\sigma_u^2}{(1-\tau_D^2)\sigma_r^4\lambda_k^2}}{p_k+\frac{\sigma_u^2}{(1-\tau_D^2)\sigma_r^2\lambda_k}}\right)\le0. \quad (31c)$$

Since the OF in (28) is non-convex, standard convex optimization algorithms are not guaranteed to solve it. Towards this end, we have the following main result.

*Lemma 1: The optimal solution of* (31) *exists if and only if*

$$\mathbf{p}^*=\arg\min_{\mathbf{p}}\{\phi(\mathbf{p})-\psi^*\varphi(\mathbf{p})\}, \quad (32)$$

*with $\psi^*$ being the unique zero of the auxiliary function $\upsilon(\psi)$, where*

$$\upsilon(\psi)=\min_{\mathbf{p}}\{\phi(\mathbf{p})-\psi\varphi(\mathbf{p})\}. \quad (33)$$

*Proof:* The proof of Lemma 1 is given in Proposition 2.1 of [47]. □

Therefore, solving problem (31) is equivalent to solving the following optimization problem:

$$\min_{\mathbf{p}}\ \phi(\mathbf{p})-\psi^*\varphi(\mathbf{p}) \quad (34a)$$

$$\text{s.t. (9c),} \quad (34b)$$

$$\gamma_0-\left(\frac{P_t}{\sigma_r^2\lambda_k}-\frac{\frac{P_t\sigma_u^2}{(1-\tau_D^2)\sigma_r^4\lambda_k^2}}{p_k+\frac{\sigma_u^2}{(1-\tau_D^2)\sigma_r^2\lambda_k}}\right)\le0. \quad (34c)$$

As a result, at the $n$th iteration of the SPCA method, we have to solve a convex problem. Let us introduce the notation of $\mathcal{G}(p_k^{(n)})=\frac{\frac{P_t\sigma_u^2}{(1-\tau_D^2)\sigma_r^4\lambda_k^2}}{p_k^{(n)}+\frac{\sigma_u^2}{(1-\tau_D^2)\sigma_r^2\lambda_k}}$ and $\mathcal{F}^{(n)}(p_k)=\gamma_0-\frac{P_t}{\sigma_r^2\lambda_k}+\mathcal{G}(p_k^{(n)})$, we have

$$\begin{aligned}\mathcal{F}^{(n)}(p_k)&=\gamma_0-\frac{P_t}{\sigma_r^2\lambda_k}+\mathcal{G}(p_k^{(n)})\\&\le\gamma_0-\frac{P_t}{\sigma_r^2\lambda_k}+\mathcal{G}(p_k^{(n-1)})\\&\quad+(p_k^{(n)}-p_k^{(n-1)})\frac{\partial\mathcal{G}(p_k^{(n)})}{\partial p_k^{(n)}}\Big|_{p_k^{(n)}=p_k^{(n-1)}}\\&=U^{(n)}(p_k),\end{aligned} \quad (35)$$

where the second inequality follows from the well-known descent lemma (see [48]). Hence, Problem (34) becomes

$$\mathrm{P_n}:\quad \min_{\mathbf{p}^T}\ \phi(\mathbf{p})-\psi^*\varphi(\mathbf{p}) \quad (36a)$$

$$\text{s.t. (9c),} \quad (36b)$$

$$U^{(n)}(p_k)\le0. \quad (36c)$$

The variable $p_k^{(n-1)}$ is a fixed parameter depending on the solution of Problem $\mathrm{P_{n-1}}$. The SPCA method is detailed in Algorithm 1. According to Sections I and II of [46], the idea of choosing an arbitrary starting point in the feasible set works well for the SPCA method. Thus, we only need to choose arbitrary initial values of $\{p_k^{(0)}\}$, $\forall k$. As shown in Algorithm 2, we employ Dinkelbach's algorithm to solve Problem (36) [47], [49]. Now each subproblem of Algorithm 2 is a convex minimization problem subject to convex constraints, which can be globally solved at each iteration. Through iterations, Algorithm 2 converges to the global optimum. Notably, Algorithm 2 can be carried out at a polynomial-time complexity due to its super-linear convergence rate [47].

Additionally, we establish a convergence result for the SPCA method in Lemma 2. Since the original problem (28) is non-convex, it is not possible to prove the convergence to

**Algorithm 1** The Framework of the Power Allocation Algorithm for Problem (28)

1: Step 0: Initialize starting point $p_k^{(0)}$ which is feasible to problem (28), and set $U^{(1)}(p_k) = \gamma_0 - \frac{P_t}{\sigma_r^2\lambda_k} + \mathcal{G}(p_k^{(0)}) + (p_k^{(1)} - p_k^{(0)})\frac{\partial \mathcal{G}(p_k^{(n)})}{\partial p_k^{(n)}}|_{p_k^{(n)}=p_k^{(0)}}$.
2: Step $n$: Compute $p_k^{(n)}$ of Problem (36);
Set $U^{(n+1)}(p_k) = \gamma_0 - \frac{P_t}{\sigma_r^2\lambda_k} + \mathcal{G}(p_k^{(n)}) + (p_k^{(n+1)} - p_k^{(n)})\frac{\partial \mathcal{G}(p_k^{(n)})}{\partial p_k^{(n)}}|_{p_k^{(n)}=p_k^{(n)}}$, and $n = n+1$.

**Algorithm 2** The Framework of the Dinkelbach's Algorithm for Problem (36)

1: Initialize $\varpi$ and $\psi_0$ with $\upsilon(\psi_0) \geq 0$, $m = 0$;
2: **while** $\upsilon(\psi_m) \geq \varpi$ **do**
3: Solve the following problem:

$$\mathbf{p}_m^{(n)*} = \arg\min_{\mathbf{p}^{(n)}} \left\{ \phi(\mathbf{p}^{(n)}) - \psi_m \varphi(\mathbf{p}^{(n)}) \right\};$$

$\psi_{m+1} = \frac{\phi(\mathbf{p}_m^{(n)*})}{\varphi(\mathbf{p}_m^{(n)*})}$;
$m = m + 1$;
4: **end while**

a global minimum, but rather to the KKT points under some regularity conditions.

*Lemma 2: Let $\{\mathbf{p}^{(n)}\}$ be the sequence generated by the SPCA method. If the sequence $\{\mathbf{p}^{(n)}\}$ converges to a regular point $\{\mathbf{p}^*\}$, then $\{\mathbf{p}^*\}$ is a KKT point of Problem* (34).

*Proof:* Please refer to Appendix B. □

### C. Joint Power and Computational Resource Optimization

Given the above results from the two subproblems, the joint power- and computational-resource optimization is formulated in Algorithm 3.

Again, the power allocation solution can be found by solving a series of convex optimization problems at a polynomial complexity. Furthermore, the subproblem of computational resources optimization is a convex one, which can be optimally solved at a polynomial complexity. In summary, the proposed alternating optimization algorithm only requires a polynomially increasing computational complexity with the problem dimension.

**Algorithm 3** Joint Power- and Computational-Resource Optimization Algorithm

1: Initialize $z = 0$, $\epsilon = 1$, and a feasible point $\mathbf{p}^{(0)}$.
2: **while** $\epsilon > 0.001$ **do**
3: $z = z + 1$;
4: Solve problem (26), and obtain $\boldsymbol{\nu}^{(z)}$.
5: Calculate $\mathbf{p}^{(z)}$ via Algorithm 2 with $\mathbf{p}^{(z-1)}$ and $\boldsymbol{\nu}^{(z)}$;
6: Calculate $\epsilon = \max_k |\frac{\nu_k^{(z)} - \nu_k^{(z-1)}}{\nu_k^{(z-1)}}|$;
7: **end while**

*Lemma 3: Algorithm 3 converges within finite iterations, since the optimal solution of Problem* (25) *monotonically decreases with the iterations.*

*Proof:* Please refer to Appendix C. □

### D. Extension to Imperfect CSIR of TN-FAN Channel

In this subsection, we extend to consider the scenario that the CSIR is imperfectly known at the FAN. Let $\hat{\mathbf{H}}$ denote the estimated TN-FAN channel CSI. Thus, the actual TN-FAN channel can be modeled as [39]

$$\mathbf{H} = \sqrt{1-\tau^2}\hat{\mathbf{H}} + \tau_S \boldsymbol{\Omega}_S, \tag{37}$$

where $\boldsymbol{\Omega}_S$ has i.i.d entries with zero mean and unit variance independent of $\hat{\mathbf{H}}$, and the parameter $\tau_S \in [0,1]$ reflects the estimation accuracy or quality of $\mathbf{H}$.

Following (13), the precoding matrix at the FAN is given by

$$\hat{\mathbf{W}} = \hat{\mathbf{G}}^\dagger \mathbf{P} \hat{\mathbf{H}}^\dagger,$$

where we have $\hat{\mathbf{G}}^\dagger = \hat{\mathbf{G}}^{\mathrm{H}}(\hat{\mathbf{G}}\hat{\mathbf{G}}^{\mathrm{H}})^{-1}$ and $\hat{\mathbf{H}}^\dagger = (\hat{\mathbf{H}}^{\mathrm{H}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^{\mathrm{H}}$. The signal vector received at the CNs can be formulated as

$$\begin{aligned}\mathbf{y}_U &= \mathbf{G}\hat{\mathbf{W}}\mathbf{H}\mathbf{x} + \mathbf{G}\hat{\mathbf{W}}\mathbf{n}_R + \mathbf{n}_U \\ &= \mathbf{G}\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger\mathbf{H}\mathbf{x} + \mathbf{G}\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger\mathbf{n}_R + \mathbf{n}_U \\ &= \sqrt{1-\tau_D^2}\sqrt{1-\tau_S^2}\mathbf{P}\mathbf{x} + \boldsymbol{\Omega}_U\mathbf{x} + \mathbf{G}\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger\mathbf{n}_R \\ &\quad + \mathbf{n}_U. \end{aligned} \tag{38}$$

where $\boldsymbol{\Omega}_U = \sqrt{1-\tau_D^2}\mathbf{P}\hat{\mathbf{H}}^\dagger\tau_S\boldsymbol{\Omega}_S + \tau_D\boldsymbol{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\sqrt{1-\tau_S^2} + \tau_D\boldsymbol{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger\tau_S\boldsymbol{\Omega}_D$ is the channel estimation error. In particular, the signal received at the $k$th CN is

$$\begin{aligned}y_k &= \sqrt{(1-\tau_D^2)(1-\tau_S^2)P_t p_k}s_k + \mathbf{g}_{U,k}\mathbf{x} \\ &\quad + \sqrt{(1-\tau_D^2)p_k}\hat{\mathbf{h}}_k^\dagger\mathbf{n}_R + \left[\tau_D\boldsymbol{\Omega}_D\hat{\mathbf{H}}_D^\dagger\mathbf{P}\hat{\mathbf{H}}_S^\dagger\right]_k\mathbf{n}_R + n_k,\end{aligned} \tag{39}$$

where $\omega_{U,k}$ is the $k$th row of $\boldsymbol{\Omega}_U$. The SINR of the $k$th data stream is characterized by

$$\hat{\gamma}_k = \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t(\omega_{U,k}\omega_{U,k}^H) + p_k(1-\tau_D^2)\sigma_r^2(\hat{\mathbf{h}}_k^\dagger(\hat{\mathbf{h}}_k^\dagger)^H) + \chi_k + \sigma_u^2}, \tag{40}$$

where $\chi_k = \tau_D^2\sigma_r^2\left[\boldsymbol{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger(\boldsymbol{\Omega}_D\hat{\mathbf{G}}^\dagger\mathbf{P}\hat{\mathbf{H}}^\dagger)^H\right]_{k,k}$. From (40), we have the following theorem.

*Theorem 2: Let $\tau_1 = (1-\tau_D^2)\tau_S^2\lambda_k$, $\tau_2 = \frac{(1-\tau_D^2)(1-\tau_S^2)\tau_D^2}{1+\tau_D^2}$ and $\tau_3 = \frac{\tau_S^2\tau_D^2\lambda_k(1-\tau_D^2)}{(1+\tau_D^2)}$. The SINR of the $k$th data stream defined in* (40) *can be expressed as*

$$\hat{\gamma}_k = \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t\alpha_\tau + \sum_{i=1}^K p_i\sigma_r^2\lambda_k(1-\tau_D^2) + \frac{1}{M}\tau_D^2 p_k\lambda_k\beta_\tau + \sigma_u^2}, \tag{41}$$

*where $\alpha_\tau = \tau_1 p_k + \frac{1}{M}\tau_2\sum_{i=1}^K p_i + \frac{1}{M}\tau_3\sum_{i=1}^K p_i$ and $\beta_\tau = \frac{1-\tau_D^2}{1+\tau_D^2}$.*

*Proof:* Please refer to Appendix D. □

Based on Theorem 2, we have the following propositions.

*Proposition 1: As the number of antennas in the FAN tends to $M \to \infty$, the SINR in* (41) *can be asymptotically expressed as*

$$\hat{\gamma}_{k,\infty} = \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t(1-\tau_D^2)\tau_S^2\lambda_k p_k + \sum_{i=1}^{K} p_i\sigma_r^2\lambda_k(1-\tau_D^2) + \sigma_u^2}. \tag{42}$$

*Proposition 2: If $K$ TNs are capable of accessing the massive MIMO-aided task offloading systems, then we have:*

$$K \leq \sqrt{\frac{(1-\tau_D^2)(1-\tau_S^2)P_r P_t - (1-\tau_D^2)\tau_S^2\lambda_k P_t P_r\gamma_0 - \sigma_u^2\gamma_0}{(1-\tau_D^2)\lambda_k\vartheta\sigma_r^2\gamma_0}}. \tag{43}$$

*Proof:* Please refer to Appendix E. □

Based on (42), the FAN power allocation problem can be expressed as

$$\min_{\mathbf{p}} \sum_{k=1}^{K} \frac{(P_t+p_k)\nu_k b_k}{\mathcal{R}_k} = \sum_{k=1}^{K} \frac{(P_t+p_k)\nu_k b_k}{\frac{B}{2}\log_2(1+\hat{\gamma}_k)} \tag{44a}$$

$$\text{s.t. (9c)}, \tag{44b}$$

$$\hat{\gamma}_k \geq \gamma_0, \quad \forall k. \tag{44c}$$

*Theorem 3: Algorithm 3 can be used to solve the robust average energy minimization problem given in* (44).

*Proof:* Please refer to Appendix F. □

## V. Simulation Results and Discussions

In this section, our simulation results characterizing the proposed task offloading strategy are presented in comparison to several baseline schemes. We consider the typical outdoor wireless propagation environment, where the channel's power angle spectrum (PAS) can be modeled by the truncated Laplacian distribution [50], [51], while the eigenvalues depend on the channel PAS, which reveals a relationship between the channel's spatial correlations and channel power distribution in the angular domain.

### A. System Parameters

Computer simulations are conducted to verify the accuracy of our analytical results, and the simulation results are obtained by averaging over 10,000 channel realizations. the channel with the parameter settings is generated using floating-point arithmetic in MATLAB. In the simulation, the channel samples are generated at a period of 0.005ms. Unless mentioned otherwise, most of the simulations obey the following scenario. There are 25 CNs with sufficient computational resources. For each TN and CN, the CPU's computational capacity is randomly selected from the set $\{0.1, 0.2, \cdots, 1.0\}$ GHz. The Random Access Memory (RAM) size is 2 GB and the local computation's energy per CPU cycle $z_i$ follows a uniform distribution in the range of $(0, 20 \times 10^{-11})$ J/cycle. For the computing task, we consider a robot mapping application similar to that in [42], [52], where the task size of any TN $k$ for the computation offloading is $a_k = 500$ KB, $\forall k \in \mathcal{S}$, the SINR threshold is 1.5 dB, and the required number of CPU cycles per bit follows the uniform distribution in $[500, 1500]$ cycles/bit.

The nodes are uniformly distributed in a square-shaped cell with a side length of $2 \times R$, where $R$ denotes the cell size. We simulate a micro-cell environment for the Non Line of Sight (NLOS) case and set the carrier frequency to $f_C = 2$ GHz. The external parameters and stochastic parameters are extracted from Chapter 3 of [53]. The FAN and TNs heights are set to be $h_{FAN} = 5$ m and $h_{TN} = 1.5$ m, respectively. The noise power is given by $\sigma_u^2 = Bk_B T_0 W$, where $B = 20$ MHz denotes the bandwidth, $k_B = 1.381 \times 10^{-23}$ represents the Boltzmann constant, $T_0 = 290$ (Kelvin) denotes the noise temperature, and $W = 9$ dB is the noise figure.

### B. Performance Evaluation

Fig. 2 shows the total energy consumption of massive MIMO-aided fog computing systems versus the number of TNs. Specifically, we compare the performance of our proposed algorithm, to pure local computing, to the maximal energy efficient task scheduling strategy (MEETS) of [9], to the full offloading strategy, to the proposed algorithm with multi-antenna relay, and to the proposed algorithm without relay under a variety of diverse conditions. Local computing and full offloading represent the scenarios that all the tasks are computed locally and remotely in the CN, respectively. We can observe from Fig. 2 that the total energy consumption increases with the number of TNs, since higher computing energy consumption and offloading energy consumption are required. It is worth noting that our proposed algorithm significantly outperforms both the pure local computing and the full offloading strategies. The objective of MEETS is to reduce the transmission energy consumption. Local computing performs better than MEETS in terms of its total energy consumption due to its reduced computational energy consumption. When the number of TNs is small, since local computing consumes much less energy than full offloading, most of the tasks are computed locally, which makes the performance of our proposed algorithm similar to that of local computing. In order to verify the performance improvement of massive MIMO, we plot the results of the proposed algorithm using a conventional relay. It can be observed that the massive MIMO scheme always performs better than the conventional multi-antenna relay. On the other hand, the transmit energy consumption is much higher than the computing energy consumption. Therefore, we can observe from the figure that the total energy consumption of the full offloading strategy is always much higher than that of the local computing strategy. However, in the regime of large task size, the delay requirement can not be guaranteed if all the tasks are computed locally. Thus, the tradeoff between local computing and full offloading strategies under the delay requirement is demonstrated quite explicitly.

We then conduct an experiment to validate the tightness of our proposed task offloading strategy. We plot the total energy consumption versus the task size for different schemes in Fig. 3, which characterizes both our proposed algorithm and the simulated optimal scheme. It can be observed that the variations of the values of the proposed algorithm and

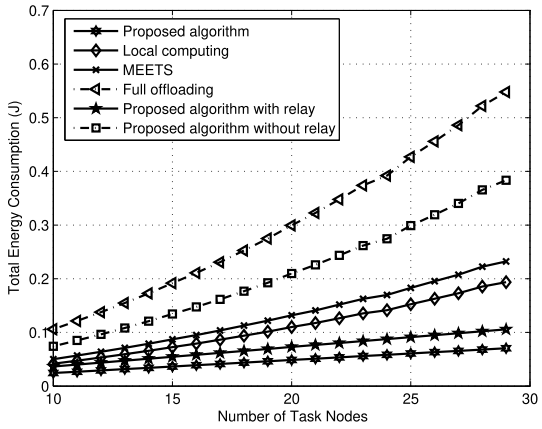



Fig. 2. Total energy consumption of the massive MIMO system versus the number of task nodes.

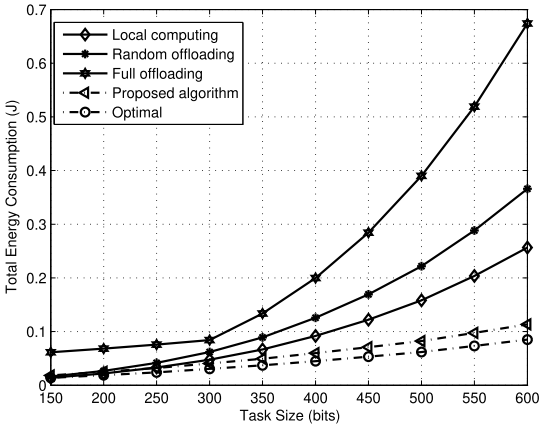



Fig. 3. Total energy consumption of the massive MIMO-enabled fog computing system versus the task size.

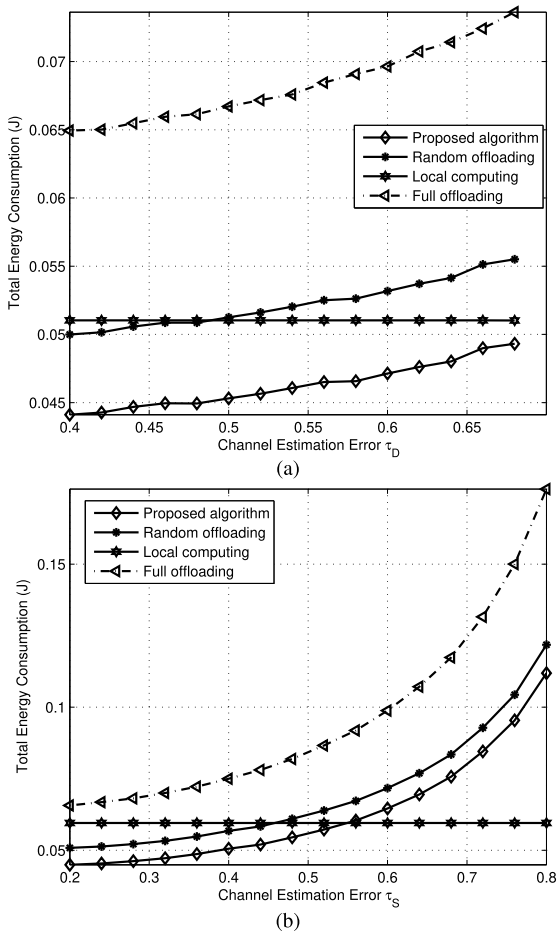



Fig. 4. Total energy consumption of massive MIMO-enabled fog computing system versus the channel estimation errors.

the simulated optimal scheme agree reasonably well. The performance of the local computing strategy approaches that of the proposed algorithm when the task size decreases and saves substantial energy over the full offloading strategy. This suggests that there exists some critical value of the task size, under which reducing the task size yields no total energy consumption reduction for the proposed algorithm compared to the local computing strategy. Additionally, we observe that the total energy consumptions of both the local computing strategy and of the proposed algorithm converge to that of the optimal solution, when the task size decreases. This is due to the fact that the local computing strategy is the most energy efficient strategy when the task size is small. Furthermore, the total energy consumption of the proposed algorithm is lower than that of the other three existing strategies when the task size is higher than 250 bits.

Fig. 4(a) and Fig. 4(b) show the total energy consumption of the massive MIMO-aided fog computing system versus the FAN-CN channel estimation quality $\tau_D$ and TN-FAN channel estimation quality $\tau_S$, respectively. With random offloading strategy, all the TNs choose the random offloading ratios. We observe from the figure that the energy consumption of the proposed algorithm is much lower than that of local computing, full offloading, and random offloading, respectively. Additionally, we observe that the total energy consumptions of the full offloading strategy, random offloading strategy, and of the proposed algorithm is increased when the channel estimation error increases. This is due to its higher transmit energy consumption. As shown in both figures, it is obvious that the total energy consumption of the local computing strategy does not vary with the channel estimation accuracy. Furthermore, it can be observed from Fig. 4(a) that there exists a crossover point between the random offloading strategy and local offloading strategy. This means that the FAN-CN channel estimation error $\tau_D$ influences the offloading decisions. This observation can be interpreted as follows: As the channel estimation error $\tau_D$ becomes large, the transmit power has to be increased to meet the SINR constraint. Hence the total energy consumption of local computing may become lower than that of offloading. Similarly, it may be observed from Fig. 4(b) that there exists a crossover point between using the local computing strategy and our proposed algorithm.

In order to further augment the interpretation of the asymptotic form of the effective SINR from Theorem 1, Fig. 5 plots the effect of different numbers of antennas on the total energy consumption, which illustrates the scenario of the total energy consumption versus the channel estimation error $\tau_D$ for different numbers of antennas. It can be observed that the total energy consumption is increased when the channel estimation error $\tau_D$ is increased regardless of the number of antennas. Furthermore, the larger the number of antennas,

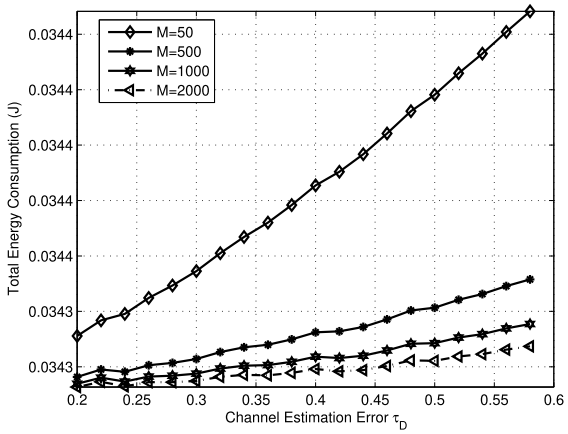



Fig. 5. Total energy consumption of massive MIMO-enabled fog computing systems versus the channel estimation error $\tau_D$.

the smaller the energy consumption reduction becomes upon having more antennas. This coincides with the analytic results of Section IV: For a large number of antennas, the SINR converges to a value that is independent of the number of antennas. These results further indicate that the total energy consumption can be reduced upon increasing the number of antennas. Therefore, we can choose the most energy efficient offloading strategy for our massive MIMO-aided fog computing system according to the asymptotic form of the effective SINR.

## VI. Conclusion

A massive MIMO-enabled task offloading framework has been proposed, where multiple TNs rely on task offloading via a massive MIMO-aided FAN to multiple CNs. We formulated an optimization problem for minimizing the total energy consumption of task offloading, in the face of imperfect CSI. In order to tackle this challenging problem, we have solved the task offloading and power allocation problem in an alternating manner. We first determined the task and computational resource allocation for a given power allocation, followed by presenting a sequential optimization framework for determining the power allocation that minimizes the total energy consumption at the TNs and FAN. Based on the task-, computational-resource, and power-allocations, we have proposed an iterative algorithm for obtaining the joint results. The simulation results showed that the proposed scheme achieves much better performance than the benchmarks. In a future work we will consider the scenario of multiple task nodes to multiple computing nodes under the proposed resource allocation framework.

## Appendix

### A. Proof of Theorem 1

For the second term on the right hand side (RHS) in (16), we expand the trace of $\tau_D \mathbf{\Omega}_D \hat{\mathbf{G}}^{\dagger} \mathbf{P} \mathbf{x}$ and obtain its power as

$$\begin{aligned} &\mathsf{E}\left[\tau_D^2 \mathbf{\Omega}_D \hat{\mathbf{G}}^{\dagger} \mathbf{P} \mathbf{x} \mathbf{x}^{\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger \mathrm{H}} \mathbf{\Omega}_D^{\mathrm{H}}\right]_{k,k} \\ &\quad = \tau_D^2 \mathrm{tr}\left(\hat{\mathbf{G}}^{\dagger} \mathbf{P} \mathbf{x} \mathbf{x}^{\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger \mathrm{H}}\right) \\ &\quad = \frac{(1-\tau_D^2)\tau_D^2 P_t \sum_{i=1} p_i}{M(1+\tau_D^2)}. \end{aligned} \tag{45}$$

Based on [54], we have the following results

$$\lim_{M\to\infty} \frac{1}{M} \mathbf{g}_i^{\mathrm{H}} \mathbf{g}_j = \begin{cases} 0, & \text{if } \ i \neq j, \\ 1, & \text{if } \ i = j. \end{cases} \tag{46}$$

Based on (46), we arrive at $\lim_{M\to\infty} \frac{1}{M} \hat{\mathbf{g}}_k^{\mathrm{H}} \hat{\mathbf{g}}_k = \frac{1+\tau_D^2}{1-\tau_D^2}$. Next we adopt the eigenvalue/eigenvector decomposition of $\mathbf{H}_k^{\mathrm{H}} \mathbf{H}_k$ to obtain

$$\mathbf{H}^{\mathrm{H}} \mathbf{H} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^{\mathrm{H}}, \tag{47}$$

where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{\mathrm{K}}\}$ and $\mathbf{Q}$ represents the nonnegative diagonal eigenvalue matrix and the unitary eigenvector matrix, respectively. Therefore, we have $\mathsf{E}\left[\mathbf{h}_k^{\mathrm{H}} \mathbf{h}_k\right] = \lambda_k$.

For the third and fourth terms on the RHS in (16), we expand the trace of $\mathbf{G} \hat{\mathbf{G}}^{\dagger} \mathbf{P} \hat{\mathbf{H}}^{\dagger} \mathbf{n}_R$ and obtain its power as follows:

$$\begin{aligned} &(1-\tau_D^2)\mathsf{E}\left[\mathbf{P} \mathbf{H}^{\dagger} \mathbf{n}_R \mathbf{n}_R^H \mathbf{H}^{\dagger H} \mathbf{P}^H\right]_{k,k} \\ &\quad + \tau_D^2 \mathsf{E}\left[\mathbf{\Omega}_D \hat{\mathbf{G}}^{\dagger} \mathbf{P} \mathbf{H}^{\dagger} \mathbf{n}_R\right]_{k,k} \\ &= (1-\tau_D^2) \sum_{i=1}^{K} p_i \mathsf{E}\left[\mathbf{h}_k^{\dagger} \mathbf{n}_R \mathbf{n}_R^{\mathrm{H}} \mathbf{h}_k^{\dagger \mathrm{H}}\right] \\ &\quad + \tau_D^2 \mathsf{E}\left[\mathbf{\Omega}_D \hat{\mathbf{G}}^{\dagger} \mathbf{P} \mathbf{H}^{\dagger} \mathbf{n}_R \mathbf{n}_R^{\mathrm{H}} \mathbf{H}^{\dagger \mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger \mathrm{H}} \mathbf{\Omega}_D^{\mathrm{H}}\right]_{k,k} \\ &= \sum_{i=1}^{K} p_i \sigma_r^2 \lambda_k (1-\tau_D^2) + \frac{\tau_D^2 \sum_{i=1}^{K} p_i \sigma_r^2 \lambda_k (1-\tau_D^2)}{M(1+\tau_D^2)}. \end{aligned} \tag{48}$$

The SINR of the $k$th data stream defined in (18) can be expressed as (49), shown at the bottom of the next page. In the large-antenna-size regime, we obtain the asymptotic form of the SINR for the $k$th data stream as

$$\gamma_{k,M\to\infty} = \gamma_{k,\infty} = \frac{(1-\tau_D^2) p_k P_t}{(1-\tau_D^2)\sigma_r^2 \lambda_k \sum_{i=1}^{K} p_i + \sigma_u^2}.$$

### B. Proof of Lemma 2

Let us assume that $\{\mathbf{p}^{(n)}\}$ is a feasible solution of Problem $P(n)$, which means that its objective function value $\phi(\mathbf{p}^{(n)}) - \psi^* \varphi(\mathbf{p}^{(n)})$ is no less than the optimal value of problem $P(n+1)$, i.e., we have $\phi(\mathbf{p}^{(n+1)}) - \psi^* \varphi(\mathbf{p}^{(n+1)}) \leq \phi(\mathbf{p}^{(n)}) - \psi^* \varphi(\mathbf{p}^{(n)})$. Additionally, since the feasible set of problem (28) is compact and nonempty, it follows that the sequence $\phi(\mathbf{p}^{(n)}) - \psi^* \varphi(\mathbf{p}^{(n)})$ is bounded, and thus has a limit.

Let us also assume that $\{\mathbf{p}^*\}$ is the convergent point. Thus $\{\mathbf{p}^*\}$ must satisfy the KKT conditions of problem (36). For any $n$, the KKT conditions are satisfied for problem $\mathrm{P_n}$, i.e., there exist nonnegative numbers $\mu_1$, $\mu_2 \in \mathbb{R}_+$ satisfying:

$$\begin{aligned} \frac{\partial\left[\phi(\mathbf{p}) - \psi^* \varphi(\mathbf{p})\right]}{\partial p_k^{(n)}} + \mu_1 + \mu_2 \frac{\partial U^{(n)}(p_k)}{\partial p_k^{(n)}} &= 0, \\ \mu_1 p_k^{(n)} &= 0, \\ \mu_2 U^{(n)}(p_k) &= 0. \end{aligned} \tag{50}$$

By denoting the limit of $\mu_1$ and $\mu_2$ by $\mu_1^*$ and $\mu_2^*$, respectively, and taking the limit $n \to \infty$ for the KKT conditions of (50), we obtain

$$\begin{aligned}\mu_2^* U(p^*) &= \mu_2 \left[\gamma_0 - \frac{P_t}{\sigma_r^2 \lambda_k} + G(p_k^{(n-1)}) \right. \\ &\quad \left. + (p_k^{(n)} - p_k^{(n-1)}) \frac{\partial G(p_k^{(n)})}{\partial p_k^{(n)}}|_{p_k^{(n)}=p_k^{(n-1)}} \right]_{n\to\infty} \\ &= \mu_2^* \left[\gamma_0 - \frac{P_t}{\sigma_r^2 \lambda_k} + G(p^*)\right] \\ &= \mu_2^* F(p^*). \end{aligned} \tag{51}$$

We finally conclude that

$$\begin{aligned}\frac{\partial \left[\phi(\mathbf{p}) - \psi^* \varphi(\mathbf{p})\right]}{\partial p^*} + \mu_1^* + \mu_2^* \frac{\partial U(p^*)}{\partial p^*} &= 0, \\ \mu_1^* p^* &= 0, \\ \mu_2^* F(p^*) &= 0, \end{aligned} \tag{52}$$

proving that $p^*$ is a KKT point of problem (34).

### C. Proof of Lemma 3

According to Algorithm 3, the $z$th iteration follows the following inequalities

$$E_{\text{total}}\left(\boldsymbol{\nu}^{(z-1)}, \mathbf{p}^{(z-1)}\right) \tag{53a}$$

$$\geq E_{\text{total}}\left(\boldsymbol{\nu}^{(z)}, \mathbf{p}^{(z-1)}\right) \tag{53b}$$

$$\geq E_{\text{total}}\left(\boldsymbol{\nu}^{(z)}, \mathbf{p}^{(z)}\right), \tag{53c}$$

where (53b) holds because problem (36) is convex and solution $\boldsymbol{\nu}^{(z)}$ represents its global optimal solution; (53c) has been proved to be valid from Lemma 2. Given (53a) and (53b), $E_{\text{total}}(\boldsymbol{\nu}, \mathbf{p})$ is reduced at each iteration. Furthermore, since $E_{\text{total}}(\boldsymbol{\nu}, \mathbf{p})$ is lower-bounded due to constraints, Algorithm 3 converges within a finite number of iterations for a given threshold.

### D. Proof of Theorem 2

Based on the following results [54]

$$\lim_{M\to\infty} \frac{1}{M} \mathbf{g}_i^{\mathrm{H}} \mathbf{g}_j = \begin{cases} 0, & \text{if} \quad i \neq j, \\ 1, & \text{if} \quad i = j, \end{cases} \tag{54}$$

we arrive at $\lim_{M\to\infty} \frac{1}{M} \hat{\mathbf{g}}_k^{\mathrm{H}} \hat{\mathbf{g}}_k = \frac{1+\tau_D^2}{1-\tau_D^2}$. We use the eigenvalue/eigenvector decomposition of $\hat{\mathbf{H}}_{S,k}^{\mathrm{H}} \hat{\mathbf{H}}_{S,k}$ to obtain

$$\hat{\mathbf{H}}^{\mathrm{H}} \hat{\mathbf{H}} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^{\mathrm{H}}, \tag{55}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_{\mathrm{K}}\}$ and $\mathbf{Q}$ represents the nonnegative diagonal eigenvalue matrix and unitary eigenvector matrix, respectively. Therefore, we have $\mathsf{E}\left[\hat{\mathbf{h}}_k^{\mathrm{H}} \hat{\mathbf{h}}_k\right] = \lambda_k$ and

$$\begin{aligned}&\mathsf{E}\left[\boldsymbol{\Omega}_U \mathbf{x}\mathbf{x}^H \boldsymbol{\Omega}_U^H\right]_{k,k} \\ &= P_t \mathsf{E}\left[\tau_S^2(1-\tau_D^2) \mathbf{P}\hat{\mathbf{H}}_S^\dagger \boldsymbol{\Omega}_S \boldsymbol{\Omega}_S^{\mathrm{H}} \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \right. \\ &\quad + \tau_D^2(1-\tau_S^2) \boldsymbol{\Omega}_D \hat{\mathbf{G}}^\dagger \mathbf{P}\mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}} \boldsymbol{\Omega}_D^{\mathrm{H}} \\ &\quad \left. + \tau_D^2 \tau_S^2 \boldsymbol{\Omega}_D \hat{\mathbf{G}}^\dagger \mathbf{P}\hat{\mathbf{H}}^\dagger \boldsymbol{\Omega}_D \boldsymbol{\Omega}_D^{\mathrm{H}} \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}} \boldsymbol{\Omega}_D^{\mathrm{H}} \right]_{k,k} \\ &= P_t \mathsf{E}\left[\tau_S^2(1-\tau_D^2) \mathbf{P}\hat{\mathbf{H}}_S^\dagger \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \right. \\ &\quad + \tau_D^2(1-\tau_S^2) \text{tr}\left(\hat{\mathbf{G}}^\dagger \mathbf{P}\mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}}\right) \mathbf{I}_{\mathrm{K}} \\ &\quad \left. + \tau_D^2 \tau_S^2 \text{tr}\left(\hat{\mathbf{G}}^\dagger \mathbf{P}\hat{\mathbf{H}}^\dagger \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}}\right) \mathbf{I}_{\mathrm{K}} \right]_{k,k} \\ &= P_t \left\{ p_k \lambda_k (1-\tau_D^2) \tau_S^2 \right. \\ &\quad + \frac{\sum_{i=1}^K p_i (1-\tau_D^2)(1-\tau_S^2) \tau_D^2}{M(1+\tau_D^2)} \\ &\quad \left. + \frac{\sum_{i=1}^K p_i \tau_S^2 \tau_D^2 \lambda_k (1-\tau_D^2)}{M(1+\tau_D^2)} \right\} \\ &= P_t \left( \tau_1 p_k + \frac{1}{M} \tau_2 \sum_{i=1}^K p_i + \frac{1}{M} \tau_3 \sum_{i=1}^K p_i \right), \end{aligned} \tag{56}$$

where $\mathsf{E}\left[\boldsymbol{\Omega}_D \mathbf{A} \boldsymbol{\Omega}_D^{\mathrm{H}}\right] = \text{tr}\left(\mathbf{A}\right) \mathbf{I}_{\mathrm{N}}$ for any $N \times N$ matrix, and $\mathsf{E}\left[\boldsymbol{\Omega}_D \boldsymbol{\Omega}_D^{\mathrm{H}}\right] = \mathbf{I}_K$ and $\mathsf{E}\left[\boldsymbol{\Omega}_S \boldsymbol{\Omega}_S^{\mathrm{H}}\right] = \mathbf{I}_M$ according to [55]. For the second term on the RHS in (38), we expand the trace of $\mathbf{G}\hat{\mathbf{G}}^\dagger \mathbf{P}\hat{\mathbf{H}}^\dagger \mathbf{n}_R$ and obtain its power as follows:

$$\begin{aligned}&\mathsf{E}\left[\mathbf{G}\hat{\mathbf{G}}^\dagger \mathbf{P}\hat{\mathbf{H}}^\dagger \mathbf{n}_R \mathbf{n}_R^{\mathrm{H}} \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}} \mathbf{G}^{\mathrm{H}}\right]_{k,k} \\ &= (1-\tau_D^2) \sum_{i=1}^K p_i \mathsf{E}\left[\hat{\mathbf{h}}_k^\dagger \mathbf{n}_R \mathbf{n}_R^{\mathrm{H}} \hat{\mathbf{h}}_k^{\dagger\mathrm{H}}\right] \\ &\quad + \tau_D^2 \mathsf{E}\left[\boldsymbol{\Omega}_D \hat{\mathbf{G}}^\dagger \mathbf{P}\hat{\mathbf{H}}^\dagger \hat{\mathbf{H}}^{\dagger\mathrm{H}} \mathbf{P}^{\mathrm{H}} \hat{\mathbf{G}}^{\dagger\mathrm{H}} \boldsymbol{\Omega}_D^{\mathrm{H}}\right]_{k,k} \\ &= \sum_{i=1}^K p_i \sigma_r^2 \lambda_k (1-\tau_D^2) + \frac{\tau_D^2 p_k \lambda_k (1-\tau_D^2)}{M(1+\tau_D^2)} \\ &= \sum_{i=1}^K p_i \sigma_r^2 \lambda_k (1-\tau_D^2) + \frac{1}{M} \tau_D^2 p_k \lambda_k \beta_\tau. \end{aligned} \tag{57}$$

Upon substituting (56) and (57) into (40), we obtain the SINR of the $k$th data stream in (41).

### E. Proof of Proposition 2

According to the Cauchy-Schwarz inequality, we have

$$\frac{K^2}{\sum_{i=1}^K \frac{1}{p_i}} \leq \sum_{i=1}^K p_i. \tag{58}$$

As a result, we have

$$\begin{aligned}\hat{\gamma}_k &\leq \hat{\gamma}_{k,\infty} \\ &= \frac{(1-\tau_D^2)(1-\tau_S^2) p_k P_t}{P_t(1-\tau_D^2)\tau_S^2 \lambda_k p_k + \sum_{i=1}^K p_i \sigma_r^2 \lambda_k (1-\tau_D^2) + \sigma_u^2} \end{aligned}$$

$$\gamma_k = \frac{(1-\tau_D^2) p_k P_t}{\frac{(1-\tau_D^2)\tau_D^2 P_t \sum_{i=1}^K p_i}{M(1+\tau_D^2)} + (1-\tau_D^2)\sigma_r^2 \lambda_k \sum_{i=1}^K p_i + \frac{(1-\tau_D^2)\tau_D^2 \sigma_r^2 \lambda_k \sum_{i=1}^K p_i}{M(1+\tau_D^2)} + \sigma_u^2} \tag{49}$$

$$\leq \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t(1-\tau_D^2)\tau_S^2\lambda_k p_k + \frac{K^2}{\sum_{i=1}^K \frac{1}{p_i}}\sigma_r^2\lambda_k(1-\tau_D^2)+\sigma_u^2}$$
$$\leq \frac{(1-\tau_D^2)(1-\tau_S^2)P_r P_t}{P_t(1-\tau_D^2)\tau_S^2\lambda_k P_r + \frac{K^2}{\sum_{i=1}^K \frac{1}{p_i}}\sigma_r^2\lambda_k(1-\tau_D^2)+\sigma_u^2}.$$

Let us specify that $\sum_{i=1}^K \frac{1}{p_i} = \frac{1}{\vartheta}$. Since $\hat{\gamma}_k \geq \gamma_0$, the above inequality can be written as

$$K \leq \sqrt{\frac{(1-\tau_D^2)(1-\tau_S^2)P_r P_t - (1-\tau_D^2)\tau_S^2\lambda_k P_t P_r \gamma_0 - \sigma_u^2\gamma_0}{(1-\tau_D^2)\lambda_k\vartheta\sigma_r^2\gamma_0}}.$$

### *F. Proof of Theorem 3*

According to (42), we have

$$\begin{aligned}\hat{\gamma}_k &\leq \hat{\gamma}_{k,\infty} \\ &= \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t(1-\tau_D^2)\tau_S^2\lambda_k p_k + \sum_{i=1}^K p_i\sigma_r^2\lambda_k(1-\tau_D^2)+\sigma_u^2} \\ &\leq \frac{(1-\tau_D^2)(1-\tau_S^2)p_k P_t}{P_t(1-\tau_D^2)\tau_S^2\lambda_k p_k + p_k\sigma_r^2\lambda_k(1-\tau_D^2)+\sigma_u^2} \\ &= \frac{(1-\tau_S^2)P_t}{P_t\tau_S^2\lambda_k + \sigma_r^2\lambda_k + \frac{\sigma_u^2}{p_k(1-\tau_D^2)}}. \end{aligned} \tag{59}$$

Therefore, $\hat{\gamma}_k$ is a concave function of $p_k$. According to the previous analytical results in subsection IV-B, the SPCA method [46] can also be applied to solve Problem (44).

## REFERENCES


[1] Y. Kawamoto, N. Yamada, H. Nishiyama, N. Kato, Y. Shimizu, and Y. Zheng, "A feedback control-based crowd dynamics management in IoT system," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1466–1476, Jul. 2017.

[2] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, and N. Kato, "A survey on network methodologies for real-time analytics of massive IoT data and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1457–1477, Apr. 2017.

[3] X. Ge, L. Pan, Q. Li, G. Mao, and S. Tu, "Multi-path cooperative communications networks for augmented and virtual reality transmission," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2345–2358, Oct. 2017.

[4] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. S. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017.

[5] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[6] C. Xu *et al.*, "Sixty years of coherent versus non-coherent tradeoffs and the road from 5G to wireless futures," *IEEE Access*, vol. 7, pp. 178246–178299, Dec. 2019.

[7] L. Pu, X. Chen, J. Xu, and X. Fu, "D2D fogging: An energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaboration," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3887–3901, Dec. 2016.

[8] X. Chen, L. Pu, L. Gao, W. Wu, and D. Wu, "Exploiting massive D2D collaboration for energy-efficient mobile edge computing," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 64–71, Aug. 2017.

[9] Y. Yang, K. Wang, G. Zhang, X. Chen, X. Luo, and M.-T. Zhou, "MEETS: Maximal energy efficient task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4076–4087, Oct. 2018.

[10] Y. Yang, S. Zhao, W. Zhang, Y. Chen, X. Luo, and J. Wang, "DEBTS: Delay energy balanced task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2094–2106, Apr. 2018.

[11] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communicatios—A key to Gigabit wireless," *Proc. IEEE*, vol. 92, no. 2, pp. 198–218, Feb. 2002.

[12] K. Wang, W. Chen, J. Li, and B. Vucetic, "Green MU-MIMO/SIMO switching for heterogeneous delay-aware services with constellation optimization," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1984–1995, May 2016.

[13] K. Wang and W. Chen, "Energy-efficient communications in MIMO systems based on adaptive packets and congestion control with delay constraints," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2169–2179, Apr. 2015.

[14] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[15] F. Rusek *et al.*, "Multiple-antenna techniques in LTE-advanced," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Oct. 2013.

[16] K. R. Liu, *Cooperative Communications and Networking*. Cambridge, U.K.: Cambridge Univ. Press., 2009.

[17] G. Amarasuriya, "Sum rate analysis for multi-user massive MIMO relay networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–7.

[18] N. Yang, M. Elkashlan, P. L. Yeoh, and J. Yuan, "Multiuser MIMO relay networks in nakagami-m fading channels," *IEEE Trans. Commun.*, vol. 60, no. 11, pp. 3298–3310, Nov. 2012.

[19] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point," in *Proc. IEEE INFOCOM*, Atlanta, GA, USA, May 2017, pp. 1–9.

[20] K. Wang, Y. Tan, Z. Shao, S. Ci, and Y. Yang, "Learning-based task offloading for delay-sensitive applications in dynamic fog networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11399–11403, Nov. 2019.

[21] K. Wang, Y. Zhou, Z. Liu, Z. Shao, X. Luo, and Y. Yang, "Online task scheduling and resource allocation for intelligent NOMA-based industrial Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 803–815, May 2020.

[22] N. Felemban *et al.*, "PicSys: Energy-efficient fast image search on distributed mobile networks," *IEEE Trans. Mobile Comput.*, early access, Dec. 31, 2019, doi: 10.1109/TMC.2019.2963150.

[23] C. Zhao, Y. Cai, A. Liu, M. Zhao, and L. Hanzo, "Mobile edge computing meets mmwave communications: Joint beamforming and resource allocation for system delay minimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2382–2396, Jan. 2020.

[24] Y. Li and W. Wang, "Can mobile cloudlets support mobile applications?" in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 1060–1068.

[25] M. Xiao, J. Wu, L. Huang, and Y. Wang, "Multi-task assignment for crowdsensing in mobile social networks," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 2227–2235.

[26] X. Chen and J. Zhang, "When D2D meets cloud: Hybrid mobile task offloadings in fog computing," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.

[27] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," Jun. 2019, *arXiv:1909.07972*. [Online]. Available: http://arxiv.org/abs/1909.07972

[28] Y. Yang, Z. Liu, X. Yang, K. Wang, X. Hong, and X. Ge, "POMT: Paired offloading of multiple tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8658–8669, Oct. 2019.

[29] C. Zhou and C.-K. Tham, "Deadline-aware peer-to-peer task offloading in stochastic mobile cloud computing systems," in *Proc. IEEE SECON*, Hong Kong, Jun. 2018, pp. 1–9.

[30] L. Chen, S. Zhou, and J. Xu, "Computation peer offloading for energy-constrained mobile edge computing in small-cell networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1619–1632, Aug. 2018.

[31] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[32] O. Y. Bursalioglu, G. Caire, R. K. Mungara, H. C. Papadopoulos, and C. Wang, "Fog massive MIMO: A user-centric seamless hotspot architecture," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 559–574, Jan. 2019.

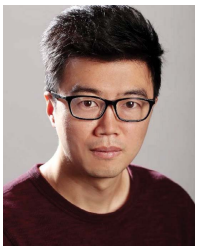
[33] H. Pirzadeh, C. Wang, and H. Papadopoulos, "Machine-learning assisted outdoor localization via sector-based fog massive MIMO," in *Proc. IEEE ICC*, Shanghai, China, Jun. 2019, pp. 1–6.

[34] D. Chen, "Low complexity power control with decentralized fog computing for distributed massive MIMO," in *Proc. IEEE WCNC*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[35] R. K. Mungara, G. Caire, O. Y. Bursalioglu, C. Wang, and H. C. Papadopoulos, "Fog massive MIMO with on-the-fly pilot contamination control," in *Proc. IEEE ISIT*, Vail, CO, USA, Jun. 2018, pp. 1–5.
[36] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals Massive MIMO*. London, U.K.: Cambridge Univ. Press, 2016.
[37] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899–3911, Jul. 2015.
[38] A. Zappone, P. Cao, and E. A. Jorswieck, "Energy efficiency optimization in relay-assisted MIMO systems with perfect and statistical CSI," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 443–457, Jan. 2014.
[39] B. Nosrat-Makouei, J. G. Andrews, and J. W. R. Heath, "MIMO interference alignment over correlated channels with imperfect CSIT," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2783–2794, Jun. 2011.
[40] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, Mar. 2017.
[41] 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, Aug. 2017.
[42] 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.
[43] G. Huertacanepa and D. Lee, "An adaptable application offloading scheme based on application behavior," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, Okinawa, Japan, Apr. 2008, pp. 387–392.
[44] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991–1995, Jun. 2012.
[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[46] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to non-convex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, no. 1, pp. 29–51, 2010.
[47] A. Roubi, "Method of centers for generalized fractional programming," *J. Optim. Theory Appl.*, vol. 107, no. 1, pp. 123–143, 2000.
[48] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.
[49] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, no. 7, pp. 492–498, Mar. 1967.
[50] K. I. Pedersen, P. E. Mogensen, and B. H. Fleury, "A stochastic model of the temporal and azimuthal dispersion seen at the base station in outdoor propagation environments," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 437–447, Mar. 2000.
[51] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM Wireless Communications With MATLAB*. Singapore: Wiley, 2010.
[52] T. Soyata, R. Muraleedharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture," in *Proc. IEEE ISCC*, Cappadocia, Turkey, Jul. 2012, pp. 59–66.
[53] R. Verdone and E. A. Zanella, *Pervasive Mobile and Ambient Wireless Communications: COST Action 2100*. Berlin, Germany: Springer, 2012.
[54] H. Q. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[55] R. Couillet and M. Debbah, *Random Matrix Methods and Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

**Kunlun Wang** (Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. From 2016 to 2017, he was with Huawei Technologies Company, Ltd., where he was involved in energy efficiency algorithm design. From 2017 to 2019, he was with the Key Laboratory of Wireless Sensor Network and Communication, SIMIT, Chinese Academy of Sciences, Shanghai, China. From 2019 to 2020, he was with the School of Information Science and Technology, ShanghaiTech University. Since 2021, he has been a Professor with the School of Communication and Electronic Engineering, East China Normal University. His current research interests include energy-efficient communications, fog computing networks, resource allocation, and optimization algorithm.

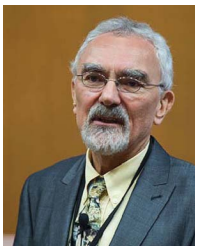
**Yong Zhou** (Member, IEEE) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From November 2015 to January 2018, he worked as a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. He is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University. His research interests include the Internet of Things, edge computing, and reconfigurable intelligent surface.



**Jun Li** (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. From January 2009 to June 2009, he worked with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell, as a Research Scientist. From June 2009 to April 2012, he was a Post-Doctoral Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow with the School of Electrical Engineering, The University of Sydney, Australia. Since June 2015, he has been a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a Visiting Professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and the industrial Internet of Things. He has coauthored more than 200 papers in IEEE journals and conferences, and holds one US patent and more than 10 Chinese patents in these areas. He was serving as an Editor of IEEE COMMUNICATION LETTERS and a TPC member for several flagship IEEE conferences. He received Exemplary Reviewer of IEEE TRANSACTIONS ON COMMUNICATIONS in 2018 and the Best Paper Award from IEEE International Conference on 5G for Future Wireless Networks in 2017.

**Long Shi** (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of New South Wales, Sydney, Australia, in 2012. From 2013 to 2016, he was a Post-Doctoral Fellow at the Institute of Network Coding, The Chinese University of Hong Kong, China. From 2014 to 2017, he was a Lecturer at the Nanjing University of Aeronautics and Astronautics, Nanjing, China. From 2017 to 2020, he was a Research Fellow with the Singapore University of Technology and Design. He is currently a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing.

**Wen Chen** (Senior Member, IEEE) is a tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is currently the Director of the Institute for Signal Processing and Systems. His research interests include multiple access, coded cooperation, and green heterogeneous networks, and has published 101 papers in IEEE journals and more than 120 papers in IEEE Conferences. He is a fellow of the Chinese Institute of Electronics and the Distinguished Lecturer of IEEE Communications Society and IEEE VTS Society. He is the Chair of IEEE VTS Shanghai Chapter, the Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.



**Lajos Hanzo** (Fellow, IEEE) received the master's and Ph.D. degrees from the Technical University (TU) of Budapest, in 1976 and 1983, respectively, the D.Sc. degree from the University of Southampton, in 2004, and the Honorary Doctorates from the TU of Budapest and the University of Edinburgh, in 2009 and 2015, respectively. He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of IEEE PRESS. He has served several terms as the Governor of both IEEE ComSoc and VTS. He has published over 1900 contributions at IEEE Xplore, 19 Wiley-IEEE Press books, and has helped the fast-track career of 123 Ph.D. students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is a fellow of the Royal Academy of Engineering (REng), IET, and EURASIP.