Task Offloading in Hybrid Intelligent Reflecting Surface and Massive MIMO Relay Networks

Kunlun Wang, *Member, IEEE*, Yong Zhou, *Member, IEEE*, Qingqing Wu, *Member, IEEE*, Wen Chen, *Senior Member, IEEE*, and Yang Yang, *Fellow, IEEE*

***Abstract*— This paper investigates the task offloading problem in a hybrid intelligent reflecting surface (IRS) and massive multiple-input multiple-output (MIMO) relay assisted fog computing system, where multiple task nodes (TNs) offload their computational tasks to computing nodes (CNs) nearby massive MIMO relay node (MRN) and fog access node (FAN) via the IRS for execution. By considering the practical imperfect channel state information (CSI) model, we formulate a joint task offloading, IRS phase shift optimization, and power allocation problem to minimize the total energy consumption. We solve the resultant non-convex optimization problem in three steps. First, we solve the IRS phase shift optimization problem with the sequential rank-one constraint relaxation (SROCR) algorithm and semi-definite relaxation (SDR) algorithm for a given power- and computational resource allocation. Then, we exploit a differential convex (DC) optimization framework to determine the power allocation decision that minimizes the total energy consumption. Given the IRS phase shifts, the computational resources, and the power allocation, we propose an alternating optimization algorithm for finding the jointly optimized results. The simulation results demonstrate the effectiveness of the proposed scheme as compared with other benchmark schemes, and the energy efficient offloading strategy for the proposed fog computing system can be chosen according to the asymptotic form of the effective signal-to-interference-plus-noise ratio (SINR).**

***Index Terms*— Task offloading, massive MIMO relay, intelligent reflecting surface, energy efficiency.**

I. Introduction

WITH the rapid development of Internet of Things (IoT), an exponentially increasing number of intelligent devices are being connected to the network [1], [2]. Meanwhile, because of the striking growth of mobile computation-intensive applications (e.g., online gaming), limited battery capacity and finite computation capacity of mobile devices pose critical challenges for the next generation wireless networks. By enabling flexible computation and communication resource sharing, fog computing (FC) as a promising technique has been proposed to offload computation intensive tasks to be executed by the nearby servers at the edge of cellular networks [3], [4]. As mobile data traffic demand is explosively increasing, massive multiple-input multiple-output (MIMO) and ultra dense heterogeneous networks (HetNets) have been proposed to enhance the system spectral efficiency (SE) and energy efficiency (EE) [5]–[8]. In addition, massive MIMO is capable of significantly improving the data rate of computational task offloading as well as the task execution efficiency. While massive MIMO has been mainly regarded as a technology for large and costly cellular base stations (BSs), the current technology trend considers higher and higher carrier frequencies and mass production for dense deployment, with corresponding decreasing size and costs. It is therefore expected that, in the near future, it will be possible to implement small and inexpensive massive MIMO nodes, each of which serving on average a relatively small number of users. However, massive MIMO systems generally require increased energy consumption and hardware cost, due to the need of installing increasingly more active antennas and/or more costly radio frequency chains operating at higher frequency bands.

A variety of sophisticated wireless communication technologies have been proposed for next generation wireless networks, including massive MIMO and intelligent reflecting

Manuscript received January 29, 2021; revised June 4, 2021, September 3, 2021, and October 20, 2021; accepted October 20, 2021. Date of publication November 2, 2021; date of current version June 10, 2022. The work of Kunlun Wang was supported by the National Natural Science Foundation of China (NSFC) under Grant 61801463. The work of Yong Zhou was supported by NSFC under Grant 62001294. The work of Qingqing Wu was supported in part by the Macau Science and Technology Development Fund, Macau, under Grant 0119/2020/A3, Grant SKL-IOTSC-2021-2023, and Grant 0108/2020/A; in part by the Guangdong NSF under Grant 2021A1515011900; and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under Grant 2021D15. The work of Wen Chen was supported in part by the National Key Project under Grant 2020YFB1807700 and Grant 2018YFB1801102, in part by NSFC under Grant 62071296, and in part by the Shanghai municipal government, the Shanghai Science and Technology Committee (STCSM) under Grant 20JC1416502. The work of Yang Yang was supported in part by the National Key Research and Development Program of China under Grant 2020YFB2104300 and in part by the NSFC Key Project Program under Grant 61932014. An earlier version of this article was presented in part at the IEEE Global Communications Conference, Madrid, Spain, December 7–11, 2021. The associate editor coordinating the review of this article and approving it for publication was Y. Liu. *(Corresponding author: Qingqing Wu.)*

Kunlun Wang is with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China (e-mail: klwang@cee.ecnu.edu.cn).

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

Qingqing Wu is with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: qingqingwu@um.edu.mo).

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

Yang Yang is with the Shanghai Institute of Fog Computing Technology (SHIFT), ShanghaiTech University, Shanghai 201210, China, also with the Research Center for Network Communication, Peng Cheng Laboratory, Shenzhen 518000, China, and also with Shenzhen Smart Cities Technology Development Group Company, Ltd., Shenzhen 518046, China (e-mail: yangyang@shanghaitech.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3122992.

Digital Object Identifier 10.1109/TWC.2021.3122992

surface (IRS). In the next generation new radio (NR) standard, reaching out beyond 6 GHz, the coverage area of each base station (BS) is significantly reduced [9], as high-frequency signals are sensitive to blockage effects [10] of obstacles, e.g., trees and buildings. On the other hand, the devices at the cell edge and/or behind line-of-sight (LoS) blockages usually suffer from low offloading rates, which increases both the latency and the energy consumption of computation offloading [11]. In order to circumvent the above limitations, IRS has been proposed as a cost-effective solution for potentially achieving high spectrum and energy efficiency via only low-cost reflecting elements [12]–[18]. Due to the combining of array aperture gain and the reflection-aided beamforming gain, IRS is capable of improving the success rate of the task offloading, hence improving the potential of FC systems. For the array aperture gain, it is generally achieved by combining both the direct and IRS-reflected signals. While for the reflection-aided beamforming gain, it is realized by controlling the phase shift of each IRS element. Given the potential gains, if the direct LoS link between the task offloading nodes and computing nodes is blocked by obstacles, the task can be offloaded via the IRS reflected link. Explicitly, reflection-based beamforming gain can be realized by jointly optimizing the IRS's phase shifts, for enhancing the offloading rate of the devices at the cell edge. Thus, IRS and massive MIMO will be the key technologies for next generation wireless networks, and FC combining next generation wireless networks holds great promise for many applications. In [19], [20], the authors have studied the impact of the IRS and the massive MIMO techniques on computational performance in a fog computing system, which have demonstrated the benefits of the IRS and massive MIMO to improve the computational offloading in the fog computing system, in comparison to the benchmark schemes. However, since most of the contributions on massive MIMO and IRS have been considered separately, there is a paucity of literature on hybrid massive MIMO and IRS. In order to exploit the benefits of massive MIMO and IRS, the multi-hop computing consists of massive MIMO relay node (MRN) computing and fog access node (FAN) computing will have new paradigm shifts for local intelligent services. This new computing paradigm will largely enhance network resilience, distributed computing and processing, and realize lower latency. Motivated by the above, we focus on investigating the role of massive MIMO and IRS in FC systems in this paper.

A. Related Works

In order to address the latency and energy efficiency issues, FC has been proposed to offload the computational tasks to be executed by the nearby servers with the powerful computing capability. Recently, task offloading has gained increasing attention in a diverse range of FC scenarios [3], [4], [21]. In particular, Chen *et al.* [22] proposed a game theoretic approach for the computation offloading decision making problem among multiple devices for mobile-edge cloud computing. As a further extension, Wang *et al.* [23] presented an alternating direction method of multipliers (ADMM)-based decentralized algorithm for computation offloading, resource allocation and content caching optimization in heterogeneous wireless cellular networks. Wang *et al.* [3] proposed a non-orthogonal multiple access (NOMA)-based FC framework for industrial Internet of things (IIoT) systems, where multiple task nodes offload their tasks via NOMA to multiple nearby helper nodes for execution. Yang *et al.* [24], [25] formulated and studied a generalized Nash equilibrium problem (GNEP) for task offloading, which effectively mapped multiple tasks or task nodes (TNs) into multiple helper nodes (HNs) to minimize every task's service delay in a distributed manner.

Leveraging a very large number of antennas at the BS, massive MIMO can significantly improve cell-throughput along with energy efficiency [26]. As expected, the integration of FC and massive MIMO can enhance the performance of task offloading in multi-user FC systems [20], [27], [28]. Resource allocation for peer offloading in fog-assisted massive MIMO networks has been widely studied. Wang *et al.* [20] proposed a massive MIMO-enabled task offloading framework, where multiple task nodes rely on task offloading via a massive MIMO-aided FAN to multiple computing nodes. Hao *et al.* [27] studied an energy-efficient multi-user computation offloading problem in massive MIMO enabled heterogeneous networks, and proposed a low-complexity alternating optimization algorithm for the joint optimization of the computational frequency of mobile devices, uplink transmit power, computational task offloading ratio and uplink transmit duration. Zeng *et al.* [28] employed massive MIMO to minimize the maximum delay for offloading and computing among all users, which requires a joint allocation of communication and computational resources. More specifically, the authors in [29] explored an edge computing-enabled cell-free massive MIMO system and analyzed the impact of the successful computation probability on the total energy consumption using queueing theory and stochastic geometry.

Although the aforementioned studies have demonstrated the benefits of massive MIMO-based FC, they have not taken into account the IRS in resource allocation and task offloading. By combining the array aperture gain and the reflection-aided beamforming gain, IRS is capable of boosting the data offloading success rate, and improving the potential of FC systems [30]. In order to exploit the benefits of IRS in wireless communications, extensive research efforts have been invested into ergodic capacity analysis [31], channel estimation [32], and practical reflection phase shift modeling [33], as well as the associated passive beamforming design in various applications [12], [34], [35]. However, the resultant optimization problems are challenging to be solved, as the optimization variables are intricately coupled. Furthermore, all the existing contributions consider the single-antenna and multiple-antenna aided IRS scenario. Owing to the rapid developments in massive MIMO FC [36], which is being increasingly adopted in IRS framework.

B. Main Contributions

In this contribution, we propose a novel hybrid FC architecture that amalgamates the benefits of both the IRS and the MRN. The nodes referred to the parlance as TNs have computationally-intensive applications to run, and hence request the multi-hop offloading of their computational tasks via the IRS to the MRN and remote FAN. After establishing the total offloading energy consumption, we formulate a joint task offloading, IRS phase shift optimization, and power allocation problem. The objective is to minimize the total energy consumption including the transmit and computation energy consumptions. Since the optimization problem is non-convex and the variables are intricately coupled, it is challenging to obtain an optimal solution with a polynomial complexity. To this end, we solve the task offloading, IRS phase shift, and power allocation problem by using the alternating optimization technique for decoupling the communication and computation designs. The main contributions of this paper are summarized as follows.

- We develop a novel hybrid massive MIMO and IRS-aided task offloading framework, where multiple TNs offload their computational tasks to MRN and FAN via an IRS. We formulate an energy minimization problem by jointly optimizing the IRS phase shift and the allocation of tasks, computational resource, and power. In particular, we consider the practical case with imperfect channel state information (CSI) and obtain the robust power allocation results.
- We partition the original optimization problem into three subproblems, namely, task and computational resource allocation subproblem, IRS phase shift optimization subproblem, and MRN power allocation subproblem. To tackle the non-convexity regarding the phase-shift vectors, we transform the phase shift optimization problem into a semidefinite programming problem (SDP), which can be solved by the sequential rank-one constraint relaxation (SROCR) algorithm and the semidefinite relaxation (SDR) algorithm. We formulate the power allocation subproblem as a non-convex optimization problem, and propose a differential convex (DC) optimization framework for the power allocation optimization. Based on the task, computational resource, IRS phase shifts, and power allocations, we propose an alternating optimization algorithm for finding the jointly optimized results. Furthermore, we prove the convergence of the proposed iterative algorithm.
- We elaborate on the impact of various parameters such as the number of IRS elements and the number of MRN antennas on the received signal-to-interference-plus-noise ratio (SINR). Our results demonstrate that the proposed algorithm achieves significant performance improvements over the benchmarks in terms of the total energy consumption.

C. Organization and Notations

The rest of the paper is organized as follows. Section II describes the system model and problem formulation. In Section III, we formulate an IRS phase shift optimization problem. In Section IV, we optimize the task offloading, computational resource allocation, IRS phase shift, and power allocation by proposing an alternating optimization algorithm for massive MIMO and IRS-aided FC networks. In Section V, we present the simulation results. Finally, the paper is concluded in Section VI.

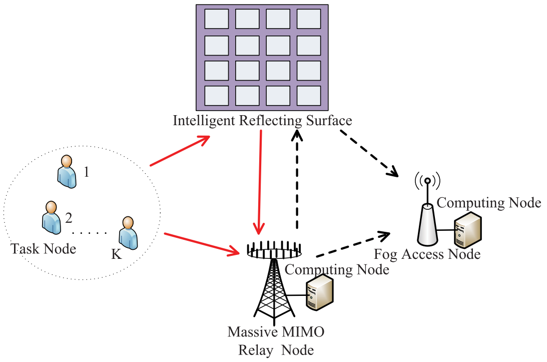

Fig. 1. Illustration of a massive MIMO and IRS-aided FC network, where K task nodes offload their tasks to a MRN and a FAN with the aid of IRS.

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. $a \sim \mathcal{CN}(0,\Gamma)$ denotes a circularly symmetric complex Gaussian variable with zero-mean and covariance Γ.

II. System Model and Problem Formulation

In this section, we first present the network model of the massive MIMO and IRS-aided FC networks, and then formulate the total energy consumption minimization problem.

A. Network Model

As shown in Fig. 1, we consider an IRS-aided FC network that operates in a time-division duplex scenario and comprises K task nodes, a decode-and-forward (DF) MRN with M antennas ($M \gg K$), an IRS with finite N reflecting elements, and an FAN with L antennas. The computing node (CN) and the FAN are assumed to be co-located and connected using a high-throughput low-latency optical fiber. Hence, the data communication between the FAN and CN is assumed to be delay-free. Similarly, the computational tasks are offloaded to the CN constituted by nearby MRN, and the CN and the MRN are also co-located and connected using high-throughput low-latency optical fiber. Then, the latency imposed by the data communication between the MRN and the CN is also deemed to be negligible. In this paper, we assume that the FAN to the TNs are assumed to be blocked by obstacles. Then, each TN can either offload its task to the MRN for computation via the IRS or to the intended FAN for computation via the IRS and MRN. Furthermore, we assume that the TNs and FAN are far from each other and there is no direct link between them.

The task offloading process consists of two hops, i.e., TNs task transmission hop and MRN task relaying hop. During the first hop, the TNs transmit their tasks to both MRN and IRS, where the latter reflects the incident tasks toward MRN (i.e., TN → IRS phase, IRS → MRN phase, and TN → MRN phase). We assume that the power of signals that are reflected by the IRS more than once is very small and can be ignored [12]. Let matrix $\mathbf{\Theta} = \mathrm{diag}(\eta_1 \mathrm{e}^{\mathrm{j}\theta_1}, \ldots, \eta_{\mathrm{N}} \mathrm{e}^{\mathrm{j}\theta_{\mathrm{N}}})$ control the reflection coefficients of the IRS elements, where $\eta_n \in [0,1]$ and $\theta_n \in [0, 2\pi)$ are the reflection amplitude and phase-shift for the nth reflecting element, respectively. During the second hop, the MRN transmits the remaining tasks to both IRS and FAN, and the IRS reflects the signals towards FAN (i.e., MRN → IRS phase, IRS → FAN phase, and MRN → FAN phase). Similarly, let diagonal matrix $\mathbf{\Phi} = \mathrm{diag}(\zeta_1 \mathrm{e}^{\mathrm{j}\phi_1}, \ldots, \zeta_{\mathrm{K}} \mathrm{e}^{\mathrm{j}\phi_{\mathrm{N}}})$ control the IRS operation during the second hop, where ζ_n and ϕ_n are the reflection amplitude and phase-shift for the nth reflecting element, respectively. It is assumed that the IRS phase shift setting is calculated at the MRN in accordance with both the channel and computing dynamics. Then, the phase shifts are sent to the IRS controller along the dedicated channel. In practice, it is costly to implement independent control of the reflection amplitude and phase shift simultaneously [12], [15]. As such, we assume the reflection amplitudes $\eta_n, \zeta_n = 1$, $\forall n$ for simplicity.

B. Channel Model

Let $\mathbf{H}_R = [\mathbf{h}_{R,1}^{\mathrm{T}}, \ldots, \mathbf{h}_{R,K}^{\mathrm{T}}] \in \mathbb{C}^{M \times K}$ denote the $M \times K$ channel coefficient matrix from K TNs to the MRN, where the kth element $\mathbf{h}_{R,k}$ denotes the channel coefficient vector between the kth TN and MRN, $k = 1, 2, \ldots, K$. Additionally, let $\mathbf{H}_S = [\mathbf{h}_{S,1}^{\mathrm{T}}, \ldots, \mathbf{h}_{S,K}^{\mathrm{T}}] \in \mathbb{C}^{N \times K}$ and $\mathbf{H}_I$ denote the $K \times N$ channel coefficient matrix from K TNs to the IRS and the $M \times N$ channel coefficient matrix from the IRS to the MRN, respectively, where the kth element $\mathbf{h}_{S,k}$ denotes the channel coefficient vector between the IRS and the kth TN, $k = 1, 2, \ldots, K$. Similarly, during the second hop task offloading, $\mathbf{H}_D$, $\mathbf{H}_{RS}$ and $\mathbf{H}_{SD}$ denote the channel coefficient matrices of the links from the MRN to the FAN, from the MRN to the IRS and from the IRS to the FAN, respectively.

In order to optimize the transmit power allocation vector $\mathbf{p} = [p_1, \ldots, p_K]$ at the MRN and IRS phase shifts, CSI is needed at the MRN, which is assumed to be perfectly available in most prior works. Perfect CSI acquisition, however, is a critical challenge due to the hardware constraint of passive RIS elements. Consequently, we assume that the CSI in each hop is imperfectly known at the MRN. Let $\hat{\mathbf{H}}_R$ denote the estimated channel coefficient matrix of the link from TNs to the MRN. In this context, the TNs to MRN channel can be modeled as [37]

$$\mathbf{H}_R = \sqrt{1 - \tau_R^2}\hat{\mathbf{H}}_R + \tau_R \mathbf{\Omega}_R, \tag{1}$$

where $\mathbf{\Omega}_R \in \mathbb{C}^{K \times M}$ has i.i.d entries with zero mean and unit variance independent of the estimated channel matrix $\hat{\mathbf{H}}_R$, and parameter $\tau_R \in [0,1]$ indicates the estimation accuracy or quality of the channel matrix $\mathbf{H}_R$.

As for the task computation, the FAN 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 FAN, the overlapping nature of the computing task makes the analysis intractable. As a result, we consider that the FAN only starts to execute the task received from the TNs after receiving all tasks. In the first hop, all TNs simultaneously transmit their symbols to the IRS and MRN in a single time slot, which is given by

$$\mathbf{x} = \sqrt{P_t}\mathbf{s}, \tag{2}$$

where P_t is the transmit power of each TN,[1] and $\mathbf{s} = [s_1, \cdots, s_K]^T$ is the transmit symbol vector with $\mathsf{E}(\mathbf{s}\mathbf{s}^{\dagger}) = \mathbf{I}_K$, and s_k is the symbol delivered from the kth TN. The signal $\mathbf{y}_R \in \mathbb{C}^{M \times 1}$ received at the MRN is

$$\mathbf{y}_R = (\mathbf{H}_R + \mathbf{H}_S \mathbf{\Theta} \mathbf{H}_I)\mathbf{x} + \mathbf{n}_R, \tag{3}$$

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

$$\mathbf{x}_R = \hat{\mathbf{W}} \mathbf{y}_R, \tag{4}$$

where $\hat{\mathbf{W}} \in \mathbb{C}^{M \times M}$ is the decoding matrix. For simplicity, we define the effective channel gain of MRN as follows:

$$\mathbf{G} = \mathbf{H}_R + \mathbf{H}_S \mathbf{\Theta} \mathbf{H}_I. \tag{5}$$

There are many ways of designing the linear receiver. Since receiver design is not the focus of this paper, we use the zero-forcing (ZF) receiver [40] in its simplicity and the assumption of ZF receiver eases the subsequent theoretical derivations to offer clear insights. Then, the MRN performs ZF precoding, the decoding matrix of the MRN can be written as

$$\hat{\mathbf{W}} = \hat{\mathbf{G}}^{\dagger}, \tag{6}$$

where $\hat{\mathbf{G}} = \hat{\mathbf{H}}_R + \hat{\mathbf{H}}_S \mathbf{\Theta} \hat{\mathbf{H}}_I$ and $\hat{\mathbf{G}}^{\dagger} = (\hat{\mathbf{G}}^{\mathrm{H}} \hat{\mathbf{G}})^{-1} \hat{\mathbf{G}}^{\mathrm{H}}$. Then, we have

$$\begin{aligned} \mathbf{G} = {} & \sqrt{1 - \tau_R^2}\hat{\mathbf{H}}_R + \tau_R \mathbf{\Omega}_R \\ & + (\sqrt{1 - \tau_S^2}\hat{\mathbf{H}}_S + \tau_S \mathbf{\Omega}_S)\mathbf{\Theta}(\sqrt{1 - \tau_I^2}\hat{\mathbf{H}}_I + \tau_I \mathbf{\Omega}_I) \\ = {} & \sqrt{1 - \tau_R^2}\sqrt{1 - \tau_S^2}\sqrt{1 - \tau_I^2}\hat{\mathbf{G}} + \tau_R \mathbf{\Omega}_R + \tau_S \tau_I \mathbf{\Omega}_S \mathbf{\Theta} \mathbf{\Omega}_I \\ & + \tau_I \sqrt{1 - \tau_S^2}\hat{\mathbf{H}}_S \mathbf{\Theta} \mathbf{\Omega}_I + \tau_S \sqrt{1 - \tau_I^2} \mathbf{\Omega}_S \mathbf{\Theta} \hat{\mathbf{H}}_I \end{aligned} \tag{7}$$

[1] We mainly focus on the optimization of power allocation for the MRN, and the assumption of fixed TN transmission power eases the subsequent theoretical derivations to offer clear insights. The authors in [38] also consider the same transmit power for each user. It is noted that for non-massive MIMO relay some existing work suggests a joint relay and user equipment (UE) power allocation for energy efficiency (EE) optimization [39]. The joint power allocation among MRN and TNs may further improve the EE performance of the massive MIMO relay systems as well, but due to limited space, this interesting work is left as a future study.

1) Received SINR at MRN: Given (1) and (6), the signal vector received at the MRN can be rewritten as

$$\begin{aligned}\mathbf{x}_R &= (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{G}\mathbf{x} + (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{n}_R\\ &= (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\Big(\sqrt{1-\tau_R^2}\sqrt{1-\tau_S^2}\sqrt{1-\tau_I^2}\hat{\mathbf{G}} + \tau_R\mathbf{\Omega}_R\\ &\quad + \tau_S\tau_I\mathbf{\Omega}_S\mathbf{\Theta}\mathbf{\Omega}_I + \tau_I\sqrt{1-\tau_S^2}\hat{\mathbf{H}}_S\mathbf{\Theta}\mathbf{\Omega}_I\\ &\quad + \tau_S\sqrt{1-\tau_I^2}\mathbf{\Omega}_S\mathbf{\Theta}\hat{\mathbf{H}}_I\Big)\mathbf{x} + (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{n}_R\\ &= \sqrt{1-\tau_R^2}\sqrt{1-\tau_S^2}\sqrt{1-\tau_I^2}\mathbf{x} + (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\left(\tau_R\mathbf{\Omega}_R\right.\\ &\quad + \tau_S\tau_I\mathbf{\Omega}_S\mathbf{\Theta}\mathbf{\Omega}_I + \tau_I\sqrt{1-\tau_S^2}\hat{\mathbf{H}}_S\mathbf{\Theta}\mathbf{\Omega}_I\\ &\quad \left. + \tau_S\sqrt{1-\tau_I^2}\mathbf{\Omega}_S\mathbf{\Theta}\hat{\mathbf{H}}_I\right)\mathbf{x} + (\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{n}_R. \qquad (8)\end{aligned}$$

According to (8), the signal received for the kth TN is

$$x_{R,k} = \sqrt{P_t(1-\tau_R^2)(1-\tau_S^2)(1-\tau_I^2)}s_k + \boldsymbol{\omega}_k\mathbf{x} + \mathbf{g}_k\mathbf{n}_R, \qquad (9)$$

where $\boldsymbol{\omega}_k$ and $\mathbf{g}_k$ are the kth rows of

$$\begin{aligned}(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\Big(&\tau_R\mathbf{\Omega}_R + \tau_S\tau_I\mathbf{\Omega}_S\mathbf{\Theta}\mathbf{\Omega}_I + \tau_I\sqrt{1-\tau_S^2}\hat{\mathbf{H}}_S\mathbf{\Theta}\mathbf{\Omega}_I\\ &+ \tau_S\sqrt{1-\tau_I^2}\mathbf{\Omega}_S\mathbf{\Theta}\hat{\mathbf{H}}_I\Big) \qquad (10)\end{aligned}$$

and $(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}$, respectively. The received SINR of the kth data stream at the MRN is given by

$$\gamma_{R,k} = \frac{(1-\tau_R^2)(1-\tau_S^2)(1-\tau_I^2)P_t}{P_t(\boldsymbol{\omega}_k\boldsymbol{\omega}_k^H) + \sigma_r^2(\mathbf{g}_k\mathbf{g}_k^H)}. \qquad (11)$$

2) Received SINR at FAN: During the second hop, the MRN transmits the decoded signal to the IRS and the FAN, where IRS reflects the incident signal towards FAN to be added constructively with the direct link from MRN. Therefore, after successfully decoding x_k at the MRN, the received signal at FAN can be given by

$$\mathbf{y}_D = (\mathbf{H}_D + \mathbf{H}_{RS}\mathbf{\Phi}\mathbf{H}_{SD})\mathbf{p}\mathbf{s} + \mathbf{n}_D, \qquad (12)$$

where $\mathbf{n}_D \in \mathbb{C}^{K\times 1}$ is the zero-mean AWGN at the FAN with a variance of $\mathsf{E}(\mathbf{n}_D\mathbf{n}_D^{\mathrm{H}}) = \sigma_r^2\mathbf{I}_K$. Similarly, we define the effective channel gain of MRN as follows:

$$\mathbf{G}_D = \mathbf{H}_D + \mathbf{H}_{RS}\mathbf{\Phi}\mathbf{H}_{SD}. \qquad (13)$$

The decoding matrix of the FAN can be written as

$$\hat{\mathbf{W}}_{\mathbf{D}} = \hat{\mathbf{G}}_D^{\dagger}, \qquad (14)$$

where $\hat{\mathbf{G}}_D = \hat{\mathbf{H}}_D + \hat{\mathbf{H}}_{RS}\mathbf{\Phi}\hat{\mathbf{H}}_{SD}$ and $\hat{\mathbf{G}}_D^{\dagger} = (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}$.

Given (12), (13) and (14), the signal received at the FAN can be derived as

$$\begin{aligned}&\mathbf{x}_D\\ &= (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{G}_D\mathbf{p}\mathbf{s} + (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{n}_D\\ &= (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\Big(\sqrt{1-\tau_D^2}\sqrt{1-\tau_{RS}^2}\sqrt{1-\tau_{SD}^2}\hat{\mathbf{G}} + \tau_D\mathbf{\Omega}_D\\ &\quad + \tau_{RS}\tau_{SD}\mathbf{\Omega}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD} + \tau_{SD}\sqrt{1-\tau_{RS}^2}\hat{\mathbf{H}}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD}\\ &\quad + \tau_{RS}\sqrt{1-\tau_{SD}^2}\mathbf{\Omega}_{RS}\mathbf{\Phi}\hat{\mathbf{H}}_{SD}\Big)\mathbf{p}\mathbf{s} + (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{n}_D\\ &= \sqrt{1-\tau_D^2}\sqrt{1-\tau_{RS}^2}\sqrt{1-\tau_{SD}^2}\mathbf{p}\mathbf{s} + (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\left(\tau_D\mathbf{\Omega}_D\right.\\ &\quad + \tau_{RS}\tau_{SD}\mathbf{\Omega}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD} + \tau_{SD}\sqrt{1-\tau_{RS}^2}\hat{\mathbf{H}}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD}\\ &\quad \left. + \tau_{RS}\sqrt{1-\tau_{SD}^2}\mathbf{\Omega}_{RS}\mathbf{\Phi}\hat{\mathbf{H}}_{SD}\right)\mathbf{p}\mathbf{s} + (\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{n}_D.\end{aligned} \qquad (15)$$

According to (15), the signal received at the FAN for the kth TN is given by

$$x_{D,k} = \sqrt{p_k(1-\tau_D^2)(1-\tau_{RS}^2)(1-\tau_{SD}^2)}s_k + \boldsymbol{\upsilon}_k\mathbf{p}\mathbf{x} + \boldsymbol{\varrho}_k\mathbf{n}_D, \qquad (16)$$

where $\boldsymbol{\upsilon}_k$ and $\boldsymbol{\varrho}_k$ are the kth rows of

$$\begin{aligned}&(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\left(\tau_D\mathbf{\Omega}_D + \tau_{RS}\tau_{SD}\mathbf{\Omega}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD}\right.\\ &\quad \left. + \tau_{SD}\sqrt{1-\tau_{RS}^2}\hat{\mathbf{H}}_{RS}\mathbf{\Phi}\mathbf{\Omega}_{SD} + \tau_{RS}\sqrt{1-\tau_{SD}^2}\mathbf{\Omega}_{RS}\mathbf{\Phi}\hat{\mathbf{H}}_{SD}\right)\end{aligned} \qquad (17)$$

and $(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}$, respectively. Accordingly, the SINR of the kth data stream received at the FAN can be given as

$$\gamma_{D,k} = \frac{(1-\tau_D^2)(1-\tau_{RS}^2)(1-\tau_{SD}^2)p_k}{p_k(\boldsymbol{\upsilon}_k\boldsymbol{\upsilon}_k^H) + \sigma_D^2(\boldsymbol{\varrho}_k\boldsymbol{\varrho}_k^H)}. \qquad (18)$$

C. Computation Model

Due to the conflict between the huge processing of complicated service and the limited amount of computing resources at MRN, a critical use case regarding the FC is the task offloading as this can save energy and/or speed up the process of computation. In general, a crucial part regarding task offloading is to decide how much and what should be offloaded [41]. In this subsection, we discuss both the MRN and the FAN computing approaches. We consider that TN k has b_k bits to be computed in a time slot of duration T. We denote the ratio of the task bits offloaded to the total task bits by ρ_k, i.e., $(1-\rho_k)b_k$ bits are offloaded to MRN for local computing and $\rho_k b_k$ bits are offloaded to FAN for remote computation.

For MRN computing, the computational power consumption for TN k can be modeled as [42]

$$P_k^L = \varrho f_k^3, \qquad (19)$$

where f_k and ϱ are the allocated CPU cycle frequency for TN k and power consumption coefficient at the MRN, respectively, which can be adjusted via the dynamic voltage and frequency scaling (DVFS) technique [41]. According to the allocated task bits for MRN computing, the computing time for TN k at the MRN is given by

$$t_k^L = \frac{\epsilon(1-\rho_k)b_k}{f_k}, \tag{20}$$

where ϵ $(\epsilon > 0)$ denotes the number of CPU cycles needed for computing each single data bit. Thus, the energy consumption of MRN computing for TN k is given by

$$E_k^L = P_k^L t_k^L = \varrho\epsilon(1-\rho_k)b_k f_k^2. \tag{21}$$

Similarly, the energy consumption of remote FAN computing for TN k is given by

$$E_k^R = P_k^R t_k^R = \varrho_R \epsilon \rho_k b_k f_{R,k}^2, \tag{22}$$

where $f_{R,k}$ and ϱ_R are the allocated CPU cycle frequency for TN k and power consumption coefficient at the FAN, respectively.

The total energy consumption E_{total} consists of the total computational energy consumptions and the total task transmit energy consumptions for the tasks. The total computational energy consumptions consist of the computational energy consumptions at the MRN and FAN, and the total task transmit energy consumptions consist of the transmit energy consumptions at the TNs and MRN, and is given by

$$E_{\text{total}} = \sum_{k=1}^{K} \left(E_k^L + E_k^R + E_k^{\text{off}} \right), \tag{23}$$

where E_k^{off} is task transmit energy consumption for TN k.

D. Problem Formulation

In this subsection, we formulate a joint IRS phase-shift matrix optimization and task-, power-, and computational-resource allocation problem for our proposed FC systems with an objective of minimizing the total energy consumption, taking into account both the communication and computational constraints. Let P_r in (24c) denote the maximum transmit power of each data stream at the MRN. To minimize E_{total} while ensuring that each TN's tasks are successfully executed within a single time slot T, the energy-efficient task offloading optimization problem is formulated as

$$\begin{aligned}
&\min_{\boldsymbol{\rho},\mathbf{p},\boldsymbol{\Theta},\boldsymbol{\Phi}} \quad E_{\text{total}} && \text{(24a)}\\
&\text{s.t.} \quad 0 \le \rho_{\text{k}} \le 1, \quad \forall \text{k}, && \text{(24b)}\\
&\qquad 0 \le p_k \le P_r, \quad \forall k, && \text{(24c)}\\
&\qquad 0 \le \theta_n \le 2\pi, 0 \le \phi_n \le 2\pi, \quad \forall n, && \text{(24d)}\\
&\qquad \gamma_{R,k}, \gamma_{D,k} \ge \gamma_0, \quad \forall k, && \text{(24e)}\\
&\qquad \sum_{k=1}^{K} f_k \le F_{RN}, && \text{(24f)}\\
&\qquad \sum_{k=1}^{K} f_{R,k} \le F_{FN}, && \text{(24g)}\\
&\qquad \frac{\epsilon(1-\rho_k)b_k}{f_k} \le T, \quad \forall k, && \text{(24h)}\\
&\qquad \frac{\epsilon\rho_k b_k}{f_{R,k}} \le T - T_0, \quad \forall k, && \text{(24i)}
\end{aligned}$$

where (24b) specifies the task offloading ratio; (24c) gives the range of the power allocation variables at the MRN; (24d) are the phase-shifts for the nth reflecting element at the IRS for the first hop transmission and second hop transmission, respectively; (24e) are the quality-of-service (QoS) constraints to ensure that both of the SINR $\gamma_{D,k}$ and $\gamma_{R,k}$ are higher than γ_0; (24f) and (24g) are imposed to ensure that the sum of computation resources allocated to all offloading TNs at the MRN and the FAN cannot exceed the total amount of computation resources F_{RN} and F_{FN} (total computational capabilities), respectively; (24h) and (24i) indicate that the delays for MRN computing and FAN computing are bounded by T and T_0, where T_0 represents the task transmission delay constraint from MRN to FAN.

Remark 1: As is shown in Problem (24)*, we have a total of four blocks of optimization variables, namely, the task offloading ratio, power allocation at the MRN, and IRS phase shifts of two hops task transmission. The optimization of the task offloading ratio is related to the computing setting, while the optimization of the power allocation at the MRN and the phase-shift matrices affects the communication design. However, Problem* (24) *is difficult to solve due to two aspects. The first one is the coupling effect between the power allocation vector* $\boldsymbol{p}$ *and the IRS phase-shift vectors* $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$*. The second one is that the objective function (OF) is non-convex with respect to the phase shifts. Thus, obtain a globally optimal solution directly is an open challenge. In this case, a locally optimal solution is provided in this paper, and we have transformations and simplifications of the original Problem* (24)*. Specifically, upon using the popular alternating optimization technique for decoupling the communications and computing designs, the optimization Problem* (24) *can be transformed to a phase shift optimization problem, a power allocation problem, and a task offloading problem, respectively. Subsequently, the optimal solutions can be provided for the power allocation* $\boldsymbol{p}$ *and for task offloading ratio* $\boldsymbol{\rho}$*, after they are decoupled from IRS phase shifts using the alternating optimization technique. In fact, alternating optimization technique is a widely applicable and empirically efficient approach for handling optimization problems involving coupled optimization variables. It has been successfully applied to several wireless communication design problems such as hybrid precoding [43], resource allocation [44], and IRS-enabled wireless communication [12], [30]. On the other hand, we transform the phase shift optimization problem into an SDP to tackle the non-convexity regarding phase shift vectors* $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$*, which can be solved by the sequential rank-one constraint relaxation (SROCR) algorithm and SDR algorithm.*

III. IRS Phase Shift Optimization

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

A. Asymptotic SINR at MRN and FAN

In the following theorem, we characterize the asymptotic property of the SINR in (11) under the massive MIMO setting, i.e., $M \to \infty$.

Theorem 1: As the number of antennas at the FAN tends to infinity, i.e., $M \to \infty$*, the effective SINR in* (11) *can be asymptotically expressed as*

$$\begin{aligned}&\gamma_{R,k,\infty}\\&=\frac{(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+N(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}{\tau_S^2(1+\tau_I^2)\sum_n(\exp(j\theta_n))^2}.\end{aligned}\tag{25}$$

Proof: Please refer to Appendix A. ■

Similarly, we characterize the asymptotic property of the SINR in (18) under the massive MIMO setting in the following theorem.

Theorem 2: As the number of antennas at the MRN tends to infinity, i.e., $M \to \infty$*, the effective SINR in* (18) *can be asymptotically expressed as (26), shown at the bottom of the page.*

Proof: Please refer to Appendix B. ■

B. Offloading Time and Energy Consumption

Given the effective SINRs in (11) and (18) of the kth data stream at the MRN and the FAN, the task offloading rate from the kth TN to the FAN is given by

$$\mathcal{R}_k=\frac{B}{2}\log_2(1+\min\{\gamma_{R,k},\gamma_{D,k}\}),\tag{27}$$

where $B/2$ is due to half-duplex working mode at the MRN. The task offloading time for the tasks allocation to MRN from TN k to the MRN is given by

$$D_{R,k}=\frac{(1-\rho_k)b_k}{B\log_2(1+\gamma_{R,k})},\quad\forall k.\tag{28}$$

For TN k, the task offloading time for the tasks allocated to FAN from the MRN to the FAN is given by

$$D_{F,k}=\frac{\rho_k b_k}{B\log_2(1+\gamma_{D,k})},\quad\forall k,\tag{29}$$

Then, based on (27), for the tasks allocated to the FAN, the offloading time from TN k to the FAN is given by

$$D_{D,k}=\frac{2\rho_k b_k}{B\log_2(1+\min\{\gamma_{R,k},\gamma_{D,k}\})}.\tag{30}$$

Next we give the total transmit energy consumption, which is given by the sum of that of the TNs and the MRN. According to the task offloading times in (28), (29) and (30), the corresponding offloading energy consumption is given by

$$\begin{aligned}&E_k^{\mathrm{off}}\\&=P_t(D_{R,k}+D_{D,k})+p_kD_{F,k}\\&=\begin{cases}\dfrac{P_t(1+\rho_k)b_k}{B\log_2(1+\gamma_{R,k})}+\dfrac{p_k\rho_kb_k}{B\log_2(1+\gamma_{D,k})}, & \gamma_{R,k}\le\gamma_{D,k},\\ \dfrac{P_t(1-\rho_k)b_k}{B\log_2(1+\gamma_{R,k})}+\dfrac{2P_t\rho_kb_k+p_k\rho_kb_k}{B\log_2(1+\gamma_{D,k})}, & \gamma_{R,k}>\gamma_{D,k}\end{cases}\end{aligned}\tag{31}$$

Given the energy consumptions of the MRN computing, the FAN computing and the task transmission in (21), (22) and (31), the total energy consumption of the massive MIMO and IRS-enabled FC system is calculated as

$$\begin{aligned}E_{\mathrm{total}}&=\sum_{k=1}^{K}\left(E_k^L+E_k^R+E_k^{\mathrm{off}}\right)\\&=\sum_{k=1}^{K}\left(\varrho\epsilon(1-\rho_k)b_kf_k^2+\varrho_R\epsilon\rho_kb_kf_{R,k}^2+E_k^{\mathrm{off}}\right).\end{aligned}\tag{32}$$

C. IRS Phase Shift Optimization

1) Optimizing Phase Shifts $\{\theta_n\}$*:* Based on (11), (31) and (32), with given fixed power allocations $\{p_k\}$ and task offloading ratios $\{\rho_k\}$, Problem (24) is formulated as

$$\min_{\mathbf{\Theta}}\quad\sum_{k=1}^{K}\frac{2(P_t+p_k)\rho_kb_k}{B\log_2(1+\gamma_{R,k})}\tag{33a}$$

$$\text{s.t.}\quad 0\le\theta_{\mathrm{n}}\le2\pi,\quad\forall\mathrm{n}.\tag{33b}$$

Let $\mathbf{v}=[v_1,\cdots,v_N]^H$, where $v_n=\exp(j\theta_n)$, $\forall n$. Denote by

$$\begin{aligned}&\varpi\\&=\frac{(1-\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\tau_S^2\tau_I^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}\\&\quad+\frac{N(1-\tau_S^2)(1+\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\tau_I^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}\\&\quad+\frac{M\tau_S^2(1-\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)(1-\tau_I^2)}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}.\end{aligned}\tag{34}$$

Then, based on the expression (11) of the received SINR $\gamma_{R,k}$ at the MRN, we obtain that the optimization variables $\{\theta_n\}$ are only related to $\boldsymbol{\omega}_k\boldsymbol{\omega}_k^H$. According to the proof of Theorem 1, we have

$$\boldsymbol{\omega}_k\boldsymbol{\omega}_k^H=\varpi\mathrm{Tr}(\mathbf{V})+|\mathbf{h}_k^{\mathrm{H}}\mathbf{h}_k|.\tag{35}$$

$$\begin{aligned}\gamma_{D,k,M\to\infty}&=\gamma_{D,k,\infty}\\&=\frac{p_k(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+p_kN\ (1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}{\tau_{RS}^2(1+\tau_{SD}^2)\sum_n(\exp(j\phi_n))^2\sum_{k=1}^{K}p_k}.\end{aligned}\tag{26}$$

Then, the phase shifts optimization problem is equivalently transformed into

$$\min_{\mathbf{V}} \quad (\boldsymbol{\omega}_k \boldsymbol{\omega}_k^H) = \varpi \mathrm{Tr}(\mathbf{V}) + |\mathbf{h}_k^{\mathrm{H}} \mathbf{h}_k| \tag{36a}$$

$$\text{s.t.} \quad \mathbf{V}_{n,n} = 1, \quad \forall n, \tag{36b}$$

$$\mathbf{V} \geq 0, \tag{36c}$$

$$\mathrm{rank}(\mathbf{V}) = 1. \tag{36d}$$

where $\mathbf{V} = \mathbf{v}\mathbf{v}^H$ and $\mathbf{h}_k$ is the kth row of $(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\tau_R \boldsymbol{\Omega}_R$. By dropping the rank-one constraint, Problem (36) is relaxed into a convex SDR problem, which can be efficiently solved by using the SDP solver in CVX [45]. It is shown that the arithmetic operation complexity of the SDP is at least $O(N^3)$ to obtain an approximate solution [46]. The optimal solution $\mathbf{V}^*$ of the SDR problem may not be rank one. If $\mathrm{rank}(\mathbf{V}^*) = 1$, then $\mathbf{V}^*$ is the optimal solution of Problem (36). Otherwise, a rank-one approximate solution needs to be extracted from $\mathbf{V}^*$ by standard rank reduction techniques, such as Gaussian randomization procedure [46]. According to [46], the Gaussian randomization procedure provides quasi-optimal bit-error-rate performance, and randomization provides an effective approximation for SDR for a sufficient (but not excessive) number of randomizations [46]. It has been shown that such an SDR approach followed by a sufficiently large number of randomizations guarantees at least a $\frac{\pi}{4}$-approximation of the optimal objective value of problems (36) and (39) [47].

On the other hand, Problem (36) can also be solved by sequential rank-one constraint relaxation (SROCR) algorithm [48]. Instead of dropping the rank-one constraint, SROCR algorithm can relax the rank-one constraint gradually such that it is easier to find a feasible solution. SROCR algorithm has been evaluated via numerical results in [48], which achieves a better performance with lower or comparable complexity compared with SDR.

2) Optimizing Phase Shifts $\{\phi_n\}$: Based on (18), (31) and (32), the second hop phase shift optimization problem is formulated as

$$\min_{\boldsymbol{\Phi}} \quad \sum_{k=1}^{K} \frac{2p_k \nu_k b_k}{B \log_2(1+\gamma_{D,k})} \tag{37a}$$

$$\text{s.t.} \quad 0 \leq \phi_{\mathrm{n}} \leq 2\pi, \forall \mathrm{n}. \tag{37b}$$

Similarly, let $\mathbf{u} = [u_1, \cdots, u_N]^H$, where $u_n = \exp(j\phi_n)$, $\forall n$.

With given power allocations $\{p_k\}$, task offloading ratios $\{v_k\}$, (18) and the proof of Theorem 2, the phase shifts optimization problem is formulated as

$$\min_{\mathbf{U}} \quad (\boldsymbol{\upsilon}_k \boldsymbol{\upsilon}_k^H) = \chi \mathrm{Tr}(\mathbf{U}) + |\mathbf{f}_k^{\mathrm{H}} \mathbf{f}_k| \tag{39a}$$

$$\text{s.t.} \quad \mathbf{U}_{n,n} = 1, \quad \forall n, \tag{39b}$$

$$\mathbf{U} \geq 0, \tag{39c}$$

$$\mathrm{rank}(\mathbf{U}) = 1. \tag{39d}$$

where χ is shown in (38) at the bottom of the page, $\mathbf{U} = \mathbf{u}\mathbf{u}^H$ and $\mathbf{f}_k$ is the kth row of $\tau_D^2(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\boldsymbol{\Omega}_D$. Problem (39) can also be solved by SROCR-based algorithm as well as relaxed into the convex SDR problem, which can be efficiently solved by using the convex optimization software as CVX [45].

IV. Joint Communication, Phase Shift and Computational Resource Optimization

In this section, we jointly optimize the task-, computation resource- and transmit power allocations to minimize the total energy consumption at the TNs and MRN. First, we solve the subproblem of task- and computational-resource allocation. Second, we solve the subproblem of power allocation at the MRN. Finally, based on the obtained results, we jointly optimize the communication, phase shift and computational resource allocation problem by conceiving an alternating optimization algorithm.

A. Task- and Computational-Resource Allocation

In this subsection, we solve the task- and computational-resource allocation subproblem under a fixed MRN power allocation. In order to solve this problem, we need to transform the non-convex optimization problem of (24) into a tractable convex optimization one.

At first, it is obvious that the OF of Problem (24) monotonically increases with f_k and $f_{R,k}$, $\forall k$. On the other hand, based on (24h) and (24i), we have the results of $f_k \geq \frac{\epsilon(1-\rho_k)b_k}{T}$ and $f_{R,k} \geq \frac{\epsilon\rho_k b_k}{T-T_0}$. Consequently, we can obtain the optimal CPU-cycle frequencies of allocated to TN k at MRN and FAN as

$$f_k^{\star} = \frac{\epsilon(1-\rho_k)b_k}{T},$$

$$f_{R,k}^{\star} = \frac{\epsilon\rho_k b_k}{T-T_0}. \tag{40}$$

Next, by substituting (40) into (24), Problem (24) is equivalently transformed into

$$\min_{\boldsymbol{\rho},\mathbf{p}} \quad \sum_{k=1}^{K} \frac{\varrho\epsilon^3(1-\rho_k)^3 b_k^3}{T^2} + \frac{\varrho_R \epsilon^3 \rho_k^3 b_k^3}{(T-T_0)^2} + E_k^{\mathrm{off}} \tag{41a}$$

$$\begin{aligned} \chi = {} & \frac{(1-\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\tau_{RS}^2\tau_{SD}^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)} \\ & + \frac{N(1-\tau_{RS}^2)(1+\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\tau_{SD}^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)} \\ & + \frac{M\tau_{RS}^2(1-\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)(1-\tau_{SD}^2)}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}. \end{aligned} \tag{38}$$

$$\begin{aligned} \text{s.t.}\quad & 0 \le \rho_{\mathrm{k}} \le 1, \quad \forall \mathrm{k}, && \text{(41b)}\\ & 0 \le p_k \le P_r, \quad \forall k, && \text{(41c)}\\ & 0 \le P_t \le P_{t_{\max}}, && \text{(41d)}\\ & \gamma_{R,k}, \gamma_{D,k} \ge \gamma_0, \quad \forall k. && \text{(41e)} \end{aligned}$$

It is should be noted that the transformed problem (41) is still non-convex. As a result, we further divide it into two tractable sub-problems of task allocation and MRN power allocation, and solve them alternately.

The task allocation subproblem with respect to the task offloading ratio is formulated as follows

$$\min_{\boldsymbol{\rho}} \quad \varphi(\boldsymbol{\rho}) = \sum_{k=1}^{K} \frac{\varrho \epsilon^3 (1-\rho_k)^3 b_k^3}{T^2} + \frac{\varrho_R \epsilon^3 \rho_k^3 b_k^3}{(T-T_0)^2} + E_k^{\mathrm{off}} \quad \text{(42a)}$$

$$\text{s.t.} \quad 0 \le \rho_{\mathrm{k}} \le 1, \quad \forall \mathrm{k}. \quad \text{(42b)}$$

As a result, we have to solve a convex problem. Towards this end, we have the following main result.

Proposition 1: The optimal task allocation ratio for each task node k of Problem (42) *is* $\rho_k^* = 1 - \frac{\Psi - \sqrt{\Lambda c_k / b_k^3 + \Lambda\Psi - \Psi c_k / b_k^3}}{\Psi - \Lambda}$, *where* $\Psi = \frac{3\varrho_R \epsilon^3}{(T - T_0^2)}$, $\Lambda = \frac{3\varrho\epsilon^3}{T^2}$ *and* $c_k = \frac{2p_k b_k}{B \log_2(1+\gamma_{D,k})}$.

Proof: Please refer to Appendix C. ■

B. MRN Power Allocation Based on DC Programming

In this subsection, we propose a DC optimization method for the MRN power allocation. By fixing the computational task offloading ratio vector $\boldsymbol{\rho}$, we only have to solve the MRN power allocation subproblem. Thus, problem (41) can be transformed into

$$\min_{\mathbf{p}} \quad \sum_{k=1}^{K} \frac{2P_t b_k \log_2(1+\gamma_{D,k}) + 2p_k \rho_k b_k \log_2(1+\gamma_{R,k})}{B \log_2(1+\gamma_{R,k}) \log_2(1+\gamma_{D,k})} \quad \text{(43a)}$$

$$\begin{aligned} \text{s.t.}\quad & (24\text{c}), (24\text{e}), && \text{(43b)}\\ & \gamma_{D,k} \ge \gamma_{R,k}, && \text{(43c)} \end{aligned}$$

where (43c) is from (31).

To begin with the problem optimization, we need to rewrite the OF of Problem (43) in the form of a single ratio as

$$\sum_{k=1}^{K} \frac{2P_t b_k \log_2(1+\gamma_{D,k}) + 2p_k \rho_k b_k \log_2(1+\gamma_{R,k})}{B \log_2(1+\gamma_{R,k}) \log_2(1+\gamma_{D,k})} = \frac{A(\mathbf{p})}{B(\mathbf{p})}, \quad \text{(44)}$$

where $A(\mathbf{p})$ and $B(\mathbf{p})$ are shown at bottom of the page, respectively.

Lemma 1: The sum-of-ratios problem (43) *is equivalent to*

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & O(\boldsymbol{p}) = 2\Upsilon \sqrt{A(\boldsymbol{p})} - \Upsilon^2 B(\boldsymbol{p}) && \text{(46a)}\\ \text{s.t.}\quad & (24\text{c}), (24\text{e}), && \text{(46b)}\\ & \gamma_{D,k} \ge \gamma_{R,k}. && \text{(46c)} \end{aligned}$$

where $\boldsymbol{\Upsilon}$ *refers to a collection of variables* $\{\Upsilon_1, \cdots, \Upsilon_K\}$.

The proof is provided in [49] and thus omitted for brevity. Due to the non-convex OF and constraints, Problem (46) is still intractable. Next we use the asymptotical form of SINR in Theorem 2 to make the problem more tractable.

C. DC Programming With the Fixed Task Offloading Ratio and IRS Phase Shift

In this subsection, the optimal solution $\mathbf{p}^*$ of problem (43) can be obtained by DC programming which is described in [50]. The DC programming is also used to solve the problems as the sum throughput maximization or the SINR max-min problem in [51], [52].

Lemma 2: Denote $\phi(\boldsymbol{p}) = 2\Upsilon\sqrt{A(\boldsymbol{p})}$ *and* $\varphi(\boldsymbol{p}) = \Upsilon^2 B(\boldsymbol{p})$, *respectively. Then, both* $\phi(\boldsymbol{p})$ *and* $\varphi(\boldsymbol{p})$ *are convex functions with respect to* $\boldsymbol{p}$ *based on Theorem 2.*

Proof: Please refer to Appendix D. ■

Proposition 2: The optimization problem (43) *is equivalent to* $\min\limits_{\boldsymbol{p}} \; (\phi(\boldsymbol{p}) - \varphi(\boldsymbol{p}))$, *which is a canonical DC programming.*

Thanks to the appropriate DC decomposition as (44), the DC programming can be exploited. Then, $\min\limits_{\mathbf{p}} \; (\phi(\mathbf{p}) - \varphi(\mathbf{p}))$ can be solved iteratively with the following convex programming. $\{\mathbf{p}_{(k+1)}\}$ at the kth iteration is generated as the optimal solution of the following convex problem:

$$\begin{aligned} \min_{\mathbf{p}} \quad & \phi(\mathbf{p}) - \varphi(\mathbf{p}_{(k)}) - \langle \bigtriangledown\varphi(\mathbf{p}_{(k)}), \mathbf{p} - \mathbf{p}_k \rangle && \text{(47a)}\\ \text{s.t.}\quad & (24\text{c}), (24\text{e}), && \text{(47b)} \end{aligned}$$

where k is the iteration index of the DC programming. The gradient of $\varphi(\mathbf{p})$ at each $\mathbf{p}$ is given by

$$\bigtriangledown\varphi(\mathbf{p}) = \left(\frac{\partial\varphi}{\partial p_1}, \frac{\partial\varphi}{\partial p_2}, \cdots, \frac{\partial\varphi}{\partial p_K} \right), \quad \text{(48)}$$

where $\frac{\partial\varphi}{\partial p_j}$, $\forall j = 1, 2, \cdots, K$ is determined by

$$\frac{\partial\varphi}{\partial p_j} = \frac{\mathcal{G} \ln 2}{1+\gamma_{D,j}} \prod_{k \ne j} \left[\frac{B}{2} \log_2(1+\gamma_{R,k}) \log_2(1+\gamma_{D,k}) \right], \quad \text{(49)}$$

where $\mathcal{G} = \frac{(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2) + N(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}{\tau_{RS}^2(1+\tau_{SD}^2)\sum_n(\exp(j\phi_n))^2 P}$.

$$\begin{aligned} A(\mathbf{p}) &= \sum_{k=1}^{K} \left\{ [2P_t b_k \log_2(1+\gamma_{D,k,\infty}) + 2p_k\rho_k b_k \log_2(1+\gamma_{R,\infty})] \prod_{l \ne k} [\log_2(1+\gamma_{D,l,\infty})] \right\}\\ B(\mathbf{p}) &= \prod_{k=1}^{K} [B \log_2(1+\gamma_{R,\infty}) \log_2(1+\gamma_{D,k,\infty})]. \end{aligned} \quad \text{(45)}$$

Since the OF in (43) is convex, we can solve Problem (43) optimally by using standard convex optimization algorithms in [45]. To this end, $\mathbf{p}_{(k+1)}$ is the optimal solution of (47) at the kth iteration, as (47) is a convex optimization problem. Then, we have the following main results.

Remark 2: Since $\varphi(\boldsymbol{p})$ is slowly sensitive to the variable $\boldsymbol{p}$, $\varphi(\boldsymbol{p})$ is well approximated by its first order approximation $\varphi(\boldsymbol{p}_{(k)}) + \langle \triangledown\varphi(\boldsymbol{p}_{(k)}),\boldsymbol{p} - \boldsymbol{p}_k\rangle$ at a fairly large neighborhood of $\boldsymbol{p}_{(k)}$. As a result, the OF $\phi(\boldsymbol{p}) - \varphi(\boldsymbol{p})$ of problem (46) *is well approximated by the convex objective* (47)*. In other words, the non-convex optimization problem* (46) *can be well approximated by the convex optimization problem* (47).

Remark 3: Since $\varphi(\boldsymbol{p})$ is convex, its gradient $\triangledown\varphi(\boldsymbol{p}_{(k)})$ is also its super-gradient [53]. Then, we have

$$\varphi(\boldsymbol{p}) \geq \varphi(\boldsymbol{p}_{(k)})+\rangle \triangledown \varphi(\boldsymbol{p}_{(k)}),\boldsymbol{p} - \boldsymbol{p}_{(k)}\rangle.$$

Therefore, program (47) *provides a well approximated upper bound minimization for the non-convex program* (46)*. To this end, upper bound minimization makes sense for intractable minimization. Since $\boldsymbol{p}_{(k)}$ is also feasible to* (47)*, we have $\phi(\boldsymbol{p}_{(k+1)}) - \varphi(\boldsymbol{p}_{(k+1)}) \leq \phi(\boldsymbol{p}_{(k)}) - [\varphi(\boldsymbol{p}_{(k)})+\rangle \triangledown \varphi(\boldsymbol{p}_{(k)}),\boldsymbol{p}_{(k+1)} - \boldsymbol{p}_{(k)}\rangle] \leq \phi(\boldsymbol{p}_{(k)}) - \varphi(\boldsymbol{p}_{(k)})$, i.e., the next solution $\boldsymbol{p}_{(k+1)}$ is always better than previous solution $\boldsymbol{p}_{(k)}$.*

Proposition 3: The sequence $\{\boldsymbol{p}_{(k)}\}$ of improved solutions always converges by Cauchy theorem, as the constraint set is compact. If $|\boldsymbol{p}_{(k)} - \boldsymbol{p}_{(k-1)}| \leq \epsilon$ or $|O(\boldsymbol{p}_{(k)}) - O(\boldsymbol{p}_{(k-1)})| \leq \epsilon$ with given threshold $\epsilon > 0$, the iterative process terminates after finite iterations.

The DC algorithm is a descent method with global convergence (i.e., from an arbitrary starting point), which does not need linesearch. The convergence of DC algorithm has been explored in [51]. The framework of the DC optimization for problem (47) is summarized in Algorithm 1, where i is the iteration index and I_τ is the maximum number of i. The complexity of Algorithm 1 mainly depends on iteratively solving Problem (47) [51]. It should be noted that the computational complexity of Problem (47) is $\mathcal{O}(K^3)$. Therefore, the computational complexity of Algorithm 1 is $\mathcal{O}(I_{\text{ite}}K^3)$, where I_{ite} denotes the number of iterations for Algorithm 1.

Algorithm 1 The Framework of the DC Optimization for Problem (47)

```
Input: $i$, $I_\tau$,
Output: $\mathbf{p}^*$
Initialize starting value $\mathbf{p}_{(0)}$ which is feasible to problem (47), set $k = 1$, and calculate $\mathbf{p}_{(1)}$, $O(\mathbf{p}_{(0)})$, and $O(\mathbf{p}_{(1)})$.
while $(|O(\mathbf{p}_{(k-1)}) - O(\mathbf{p}_{(k)})| > \epsilon)$ and $(i < I_\tau)$ do
    $k = k + 1$;
    For $\mathbf{p}_{(k-1)}$, solve (47) to obtain $\mathbf{p}_{(k)}$ by convex optimization.
    Calculate $\mathbf{p}_{(k)}$;
end while
return $\mathbf{p}_* = \mathbf{p}_{(k)}$.
```

D. Joint Power and Phase Shift Optimization

With the results from the two subproblems in place, the joint power and phase shift optimization is formulated in Algorithm 2.

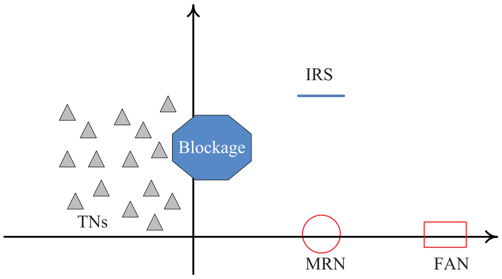

Fig. 2. Simulation setup for a massive MIMO and IRS-aided FC system, which consists of 15 TNs.

As pointed out in previous subsection, the subproblem of power allocation is given by solving a series of convex optimization problems at a polynomial complexity. Furthermore, as the subproblem of phase shift optimization is transformed into the SDP problem, it can be solved at a polynomial complexity. In all, our proposed alternating optimization algorithm only requires a polynomially computational complexity dominated by the problem dimension.

Algorithm 2 Joint Power-, Computational-Resource and Phase Shift Optimization Algorithm

```
Initialize $z = 0$, $\epsilon = 1$, and feasible points $\mathbf{p}^{(0)}$ and $\boldsymbol{\rho}^{(0)}$.
while $\epsilon > 0.001$ do
    $z = z + 1$;
    Calculate $\boldsymbol{\Theta}^{(z)}$ and $\boldsymbol{\Phi}^{(z)}$ via optimized Problems (36) and (39), respectively, with $\mathbf{p}^{(z-1)}$ and $\boldsymbol{\rho}^{(z-1)}$;
    Solve problem (42), and obtain $\boldsymbol{\rho}^{(z)}$.
    Calculate $\mathbf{p}^{(z)}$ via Algorithm 1 with $\mathbf{p}^{(z-1)}$, $\boldsymbol{\Theta}^{(z)}$, $\boldsymbol{\Phi}^{(z)}$ and $\boldsymbol{\rho}^{(z)}$;
    Calculate $\epsilon = \max_k |\frac{E_{\text{total}}^{(z)} - E_{\text{total}}^{(z-1)}}{E_{\text{total}}^{(z-1)}}|$;
end while
```

Proposition 4: Algorithm 2 is guaranteed to converge within finite iterations.

Proof: Please refer to Appendix E. ■

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.

A. System Parameters

Unless specified otherwise, the simulation parameters are set as follows. There are 15 TNs for task offloading, which are uniformly scattered within a 100 m × 100 m square area in the half left-hand side. The MRN, IRS and FAN are located at $(100, 0)$, (100,100) and (200,0) respectively, in the half right-hand side. The schematic system model for the simulated massive MIMO and IRS-aided FC network is shown in Fig. 2. For the communications channel, we consider both the small scale fading and the large scale path loss. Without loss of generality, the small scale fading is i.i.d. and obeys the complex Gaussian distribution associated with zero mean and

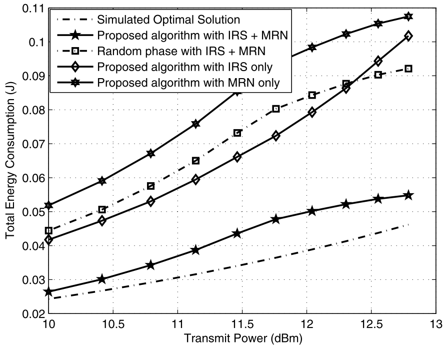

Fig. 3. Total energy consumption of massive MIMO and IRS-aided fog computing systems versus the transmit power.

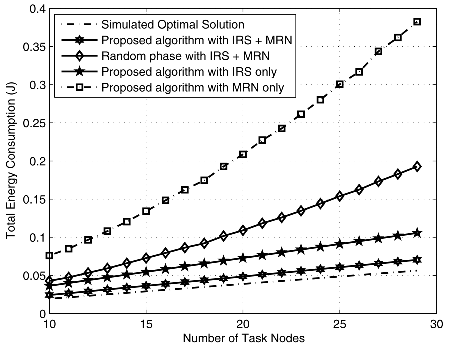

Fig. 4. Total energy consumption of massive MIMO and IRS-aided fog computing systems versus the number of task nodes.

unit variance, while the path loss in dB is given by

$$PL = PL_0 - 10\alpha \log_{10}(\frac{d}{d_0}), \tag{50}$$

where PL_0 is the path loss at the reference distance d_0; d and α represent the distance of the communications link and its path loss exponent, respectively. The transmission bandwidth of the system is $B = 20$ MHz. The variance of the AWGN is set to be 10^{-9} W. For each CN, the CPU's computational capacity is randomly selected from the set $\{0.1, 0.2, \cdots, 1.0\}$ GHz. For the computing task, we consider a robot mapping application similar to that in [22], [54], 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. Other system parameters are set as follows: $PL_0 = -30$dB, $d_0 = 50$m, and $M = 100$ (if not specified otherwise).

B. Performance Evaluation

Fig. 3 shows the total energy consumption of the massive MIMO and IRS-aided FC systems versus the transmit power of TNs. Specifically, we compare the performance of our proposed algorithm, simulated optimal solution, our proposed algorithm with IRS only, and our proposed algorithm with random phase under different conditions. The simulated optimal solution is obtained by averaging values from 2000 simulations, which is from the actual SINR at MRN and FAN, not the asymptotic SINR at MRN and FAN from Theorem 1 and 2. The benchmark of random phase of IRS has been proposed in [12], [30]. In random phase scheme, task offloading ratio and power allocation can be optimized relying on Proposition 1 and Algorithm 1, respectively, while skipping the step of designing the IRS phase shift, which is randomly set obeying the uniform distribution in the range of $[0, 2\pi)$. This figure shows the energy consumptions of our proposed algorithm and simulated optimal solution. It can be observed that the variations of analytical and simulation results agree reasonably well. We can observe from Fig. 3 that the total energy consumption increases with the transmit power of TNs, as higher offloading energy consumptions are required. As expected, our proposed algorithm always achieves better performance than that of the random phase strategy and the proposed algorithm with IRS only for different transmit power consumptions. It is worth noting that the proposed algorithm with IRS only performs better than that of the random phase strategy when the transmit power is small, as the MRN has larger energy consumption if the transmit SNR is smaller. This result makes the total energy consumption of the proposed algorithm with IRS only smaller than that of random phase with both IRS and MRN. In particular, the energy consumption of the proposed algorithm with MRN only is larger than that of the proposed algorithm with IRS. This is because by applying the proposed optimization framework with IRS, the SINRs of the TNs can be improved by providing them with additional array aperture gain and the reflection-aided beamforming gain, which leads to an improvement of the system performance.

Fig. 4 shows that the total energy consumption of the massive MIMO and IRS-aided FC systems versus the number of TNs. It is obvious that if the number of TNs is larger, then the energy consumption is higher. As expected, the proposed algorithm with MRN and IRS offers indeed the best performance than those of the other strategies, and the variations of the proposed algorithm with MRN and IRS and the simulation results agree reasonably well. Furthermore, the energy consumption of the proposed algorithm with MRN only is larger than that of the proposed algorithm with IRS. Based on (32), we can get the result that the total power consumption is larger when MRN forwards the task. It can also be observed from the figure that the energy consumption of the random phase with IRS and MRN is larger than that of the proposed algorithm with IRS only, implying that it is more beneficial for improving the system performance by deploying an IRS in the MRN-aided FC systems.

In Fig. 5, we plot the total energy consumption versus the number of IRS elements. It can be observed that the performance of the proposed algorithm with IRS only approaches converges to that of the proposed algorithm with IRS and MRN when the number of IRS elements increases and saves substantial energy over the random phase strategy. This suggests that there exists some critical value of the number of IRS elements, under which placing MRN yields no total energy consumption reduction for the proposed algorithm compared to the IRS only strategy. As shown in Fig. 5, we can observe that the SROCR algorithm performs better than SDR. This is

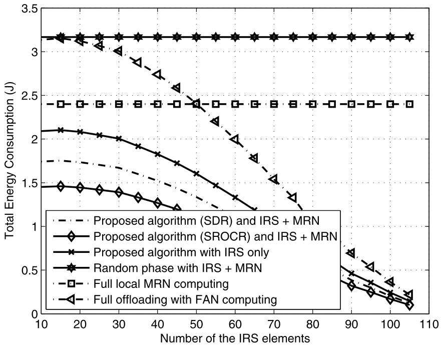

Fig. 5. Total energy consumption of massive MIMO and IRS-aided fog computing systems versus the number of IRS elements.

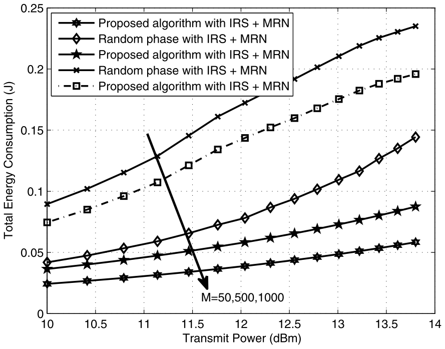

Fig. 6. Total energy consumption of massive MIMO and IRS-enabled fog computing systems versus the transmit power P_t.

expected since SROCR algorithm gradually relaxes the rank-one constraint, then it is easier to find a feasible solution. However, SDR approach only provides an approximate solution. Additionally, it is obvious that the total energy consumption of the full local MRN computing strategy does not vary with the number of IRS elements. Furthermore, the total energy consumption of the proposed algorithm is lower than that of the full offloading with FAN computing. Another interesting observation is that the total energy consumption of the full offloading with FAN computing strategy is lower than that of the full local MRN computing strategy. These further indicate that the computation energy consumption dominates the total energy consumption in the massive MIMO and IRS-aided fog computing system.

Fig. 6 shows the total energy consumption of the massive MIMO and IRS-aided FC system versus the transmit power P_t. The proposed algorithm with MRN and IRS shows a significant gain, under the same transmit power consumption, compared to the random phase strategy comprises MRN and IRS. Additionally, we observe that the total energy consumptions of proposed algorithm with MRN and IRS are decreased when the number of MRN antennas is increased, mainly due to less transmit energy consumption. Furthermore, it can be observed from Fig. 6 that the gap of the energy consumption is decreasing when increasing the number of MRN antennas. This coincides with the analytic results of Theorems 1 and 2: For a large M regime, the SINR converges to a value that is independent of the number of antennas. Therefore, we can choose the most energy efficient offloading strategy for our massive MIMO and IRS-aided fog computing system according to the asymptotic form of the effective SINR.

VI. Conclusion

In this paper, we proposed a massive MIMO and IRS-enabled task offloading framework, where multiple TNs offload their computational tasks to the MRN and FAN via IRS. We formulated an optimization problem for minimizing the total energy consumption of task offloading, considering imperfect CSI. In order to tackle this challenging problem, we solved the IRS phase shifts, task offloading and power allocation problem in an alternating manner. We first determined the IRS phases shifts for a given power allocation, followed by presenting a DC optimization framework for determining the power allocation that minimizes the total energy consumption. Based on the IRS phase shifts, task-, computational-resource, and power-allocations, we proposed an efficient alternating optimization algorithm. The simulation results showed that the proposed scheme achieves much better performance than the benchmarks, and the most energy efficient offloading strategy for our proposed massive MIMO and IRS-aided fog computing system can be chosen according to the asymptotic form of the effective SINR.

Appendix

A. Proof of Theorem 1

Based on [55], we have

$$\lim_{N\to\infty}\frac{1}{N}\mathbf{h}_{S,i}^{\mathrm{H}}\mathbf{h}_{S,j}=\begin{cases}0, & \text{if} \quad i\neq j,\\ 1, & \text{if} \quad i=j.\end{cases}\tag{51}$$

Then, we derive $\lim_{N\to\infty}\frac{1}{N}\hat{\mathbf{h}}_{S,k}^{\mathrm{H}}\hat{\mathbf{h}}_{S,k}=\frac{1+\tau_S^2}{1-\tau_S^2}$ with (51). Similarly, we have $\lim_{M\to\infty}\frac{1}{M}\hat{\mathbf{h}}_{R,k}^{\mathrm{H}}\hat{\mathbf{h}}_{R,k}=\frac{1+\tau_R^2}{1-\tau_R^2}$.

For the second term on the right hand side (RHS) in (8), we expand the trace of (10) and obtain its power as

$$\begin{aligned}&\tau_R^2\mathsf{E}\left[(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{\Omega}_R\right]_{k,k}\\ &\quad+\tau_S^2\tau_I^2\mathsf{E}\left[(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{\Omega}_S\mathbf{\Theta}\mathbf{\Omega}_I\right]_{k,k}\\ &\quad+\tau_I^2(1-\tau_S^2)\mathsf{E}\left[(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{H}}_S\mathbf{\Theta}\mathbf{\Omega}_I\right]_{k,k}\\ &\quad+\tau_S^2(1-\tau_I^2)\mathsf{E}\left[(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{\Omega}_S\mathbf{\Theta}\hat{\mathbf{H}}_I\right]_{k,k}\\ &=\frac{(1-\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\tau_R^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}\\ &\quad+\frac{(1-\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\tau_S^2\tau_I^2\sum_n(\exp(j\theta_n))^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}\\ &\quad+\frac{N(1-\tau_S^2)(1+\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\tau_I^2\sum_n(\exp(j\theta_n))^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}\\ &\quad+\frac{M\tau_S^2(1-\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)(1-\tau_I^2)\sum_n(\exp(j\theta_n))^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}.\end{aligned}\tag{52}$$

For the third term on the RHS in (8), we expand the trace of $(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{n}_R$ and obtain its power as follows:

$$\begin{aligned}&\mathsf{E}\left[(\hat{\mathbf{G}}^{\mathrm{H}}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^{\mathrm{H}}\mathbf{n}_R\mathbf{n}_R^{\mathrm{H}}\hat{\mathbf{G}}(\hat{\mathbf{G}}\hat{\mathbf{G}}^{\mathrm{H}})^{-1}\right]_{k,k}\\ &=\frac{(1-\tau_S^2)(1-\tau_I^2)(1-\tau_R^2)\sigma_R^2}{M(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+MN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}.\end{aligned}\tag{53}$$

We can obtain the SINR of the kth data stream as (54) according to (11), shown at the bottom of the page. In the large M regime, we obtain the asymptotic form of the SINR for the kth data stream as

$$\begin{aligned}&\gamma_{R,k,M\rightarrow\infty}\\ &=\gamma_{R,k,\infty}\\ &=\frac{(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+N(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}{\tau_S^2(1+\tau_I^2)\sum_n(\exp(j\theta_n))^2}.\end{aligned}\tag{55}$$

B. Proof of Theorem 2

Based on (51), we arrive at $\lim_{M\rightarrow\infty}\frac{1}{M}\hat{\mathbf{h}}_{D,k}^{\mathrm{H}}\hat{\mathbf{h}}_{D,k}=\frac{1+\tau_D^2}{1-\tau_D^2}$ and $\lim_{N\rightarrow\infty}\frac{1}{N}\hat{\mathbf{h}}_{SD,k}^{\mathrm{H}}\hat{\mathbf{h}}_{SD,k}=\frac{1+\tau_{SD}^2}{1-\tau_{SD}^2}$. For the second term on the RHS in (8), we expand the trace of (17) and obtain its power as (56), shown at the bottom of the page.

For the third term on the RHS in (15), we expand the trace of $(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{n}_D$ and obtain its power as follows in (57), shown at the bottom of the page.

We can obtain the SINR of the kth data stream as (58) according to (18), shown at the bottom of the page. In the large M regime, we obtain the asymptotic form of the SINR for the kth data stream as (59), shown at the bottom of the page.

C. Proof of Proposition 1

When $\gamma_{R,k}\leq\gamma_{D,k}$, by taking the derivative of the objective function with respect to ν_k, we have

$$\begin{aligned}\frac{\partial\varphi(\boldsymbol{\rho})}{\partial\rho_k}&=\frac{-3\varrho\epsilon^3(1-\rho_k)^2b_k^3}{T^2}+\frac{3\varrho_R\epsilon^3\rho_k^2b_k^3}{(T-T_0)^2}\\ &\quad+\frac{2p_kb_k}{B\log_2(1+\gamma_{D,k})}=0.\end{aligned}\tag{60}$$

$$\gamma_{R,k}=\frac{P_tM(1+\tau_R^2)(1-\tau_I^2)(1-\tau_S^2)+P_tMN(1+\tau_S^2)(1+\tau_I^2)(1-\tau_R^2)}{(\tau_R^2+\tau_S^2\tau_I^2\sum_n(\exp(j\theta_n))^2+N(1+\tau_S^2)\tau_I^2\sum_n(\exp(j\theta_n))^2+M\tau_S^2(1+\tau_I^2)\sum_n(\exp(j\theta_n))^2)P_t+\sigma_R^2}\tag{54}$$

$$\begin{aligned}&\tau_D^2\mathsf{E}\left[(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\boldsymbol{\Omega}_D\right]_{k,k}+\tau_{RS}^2\tau_{SD}^2\mathsf{E}\left[(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\boldsymbol{\Omega}_{RS}\boldsymbol{\Phi}\boldsymbol{\Omega}_{SD}\right]_{k,k}\\ &\quad+\tau_{SD}^2(1-\tau_{RS}^2)\mathsf{E}\left[(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{H}}_{RS}\boldsymbol{\Phi}\boldsymbol{\Omega}_{SD}\right]_{k,k}\\ &\quad+\tau_{RS}^2(1-\tau_{SD}^2)\mathsf{E}\left[(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\boldsymbol{\Omega}_{RS}\boldsymbol{\Phi}\hat{\mathbf{H}}_{SD}\right]_{k,k}\\ &=\frac{(1-\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\tau_D^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}\\ &\quad+\frac{(1-\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\tau_{RS}^2\tau_{SD}^2\sum_n(\exp(j\phi_n))^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)}\\ &\quad+\frac{N(1-\tau_{RS}^2)(1+\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\tau_{SD}^2\sum_n(\exp(j\phi_n))^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)}\\ &\quad+\frac{M\tau_{RS}^2(1-\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_{RS}^2)(1-\tau_{SD}^2)\sum_n(\exp(j\phi_n))^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}.\end{aligned}\tag{56}$$

$$\mathsf{E}\left[(\hat{\mathbf{G}}_D^{\mathrm{H}}\hat{\mathbf{G}}_D)^{-1}\hat{\mathbf{G}}_D^{\mathrm{H}}\mathbf{n}_D\mathbf{n}_D^{\mathrm{H}}\hat{\mathbf{G}}_D(\hat{\mathbf{G}}_D\hat{\mathbf{G}}_D^{\mathrm{H}})^{-1}\right]_{k,k}=\frac{(1-\tau_{RS}^2)(1-\tau_{SD}^2)(1-\tau_D^2)\sigma_D^2}{M(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+MN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}.\tag{57}$$

$$\gamma_{D,k}=\frac{p_kM(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+p_kMN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}{(\tau_D^2+\tau_{RS}^2\tau_{SD}^2\sum_n(\exp(j\phi_n))^2+N(1+\tau_{RS}^2)\tau_{SD}^2\sum_n(\exp(j\phi_n))^2+M\tau_{RS}^2(1+\tau_{SD}^2)\sum_n(\exp(j\phi_n))^2)\sum_{k=1}^{K}p_k+\sigma_D^2}\tag{58}$$

$$\begin{aligned}\gamma_{D,k,M\rightarrow\infty}&=\gamma_{D,k,\infty}\\ &=\frac{p_k(1+\tau_D^2)(1-\tau_{SD}^2)(1-\tau_{RS}^2)+p_kN(1+\tau_{RS}^2)(1+\tau_{SD}^2)(1-\tau_D^2)}{\tau_{RS}^2(1+\tau_{SD}^2)\sum_n(\exp(j\phi_n))^2\sum_{k=1}^{K}p_k}.\end{aligned}\tag{59}$$

On the other hand, when $\gamma_{R,k} > \gamma_{D,k}$, we have

$$\frac{\partial \varphi(\boldsymbol{\rho})}{\partial \rho_k} = \frac{-3\varrho\epsilon^3(1-\rho_k)^2 b_k^3}{T^2} + \frac{3\varrho_R\epsilon^3\rho_k^2 b_k^3}{(T-T_0)^2} + \frac{2(P_t+p_k)b_k}{B\log_2(1+\gamma_{D,k})} - \frac{2P_t b_k}{B\log_2(1+\gamma_{R,k})} = 0. \quad (61)$$

Denote by $\Psi = \frac{3\varrho_R\epsilon^3}{(T-T_0^2)}$, $\Lambda = \frac{3\varrho\epsilon^3}{T^2}$, $c_k = \frac{2p_k b_k}{B\log_2(1+\gamma_{D,k})}$ and $\hat{c}_k = \frac{2(P_t+p_k)b_k}{B\log_2(1+\gamma_{D,k})} - \frac{2P_t b_k}{B\log_2(1+\gamma_{R,k})}$, we have

$$\begin{aligned} -\Lambda(1-\rho_k^2)b_k^3 + \Psi\rho_k^2 b_k^3 + c_k &= 0, \quad \gamma_{R,k} \le \gamma_{D,k}, \\ -\Lambda(1-\rho_k^2)b_k^3 + \Psi\rho_k^2 b_k^3 + \hat{c}_k &= 0, \quad \gamma_{R,k} > \gamma_{D,k}. \end{aligned} \quad (62)$$

According to (62), we arrive at the optimal solution

$$\rho_k^* = \begin{cases} 1 - \dfrac{\Psi - \sqrt{\Lambda c_k/b_k^3 + \Lambda\Psi - \Psi c_k/b_k^3}}{\Psi - \Lambda}, & \gamma_{R,k} \le \gamma_{D,k}, \\ 1 - \dfrac{\Psi - \sqrt{\Lambda \hat{c}_k/b_k^3 + \Lambda\Psi - \Psi \hat{c}_k/b_k^3}}{\Psi - \Lambda}, & \gamma_{R,k} > \gamma_{D,k}. \end{cases} \quad (63)$$

D. Proof of Lemma 2

According to (26), $\gamma_{D,k,\infty}(\mathbf{p})$ is linear (hence convex), $A(\mathbf{p})$ is convex. Since the product of increasing functions is still an increasing function, $B(\mathbf{p})$ is an increasing function. $\log_2(1+\gamma_{D,k,\infty})$ are convex positive functions on a convex set, then the their geometric mean $\left[\prod_{k=1}^{K}\log_2(1+\gamma_{D,k,\infty})\right]^{1/K}$ is a convex function on the convex set [53]. Since $0 \le p_k, \forall k$, $B(\mathbf{p})$ is convex. Hence, $A(\mathbf{p})$ and $B(\mathbf{p})$ are convex functions.

When $\boldsymbol{\Upsilon}$ is fixed, due to the convexity of $A(\mathbf{p})$, the convexity of $B(\mathbf{p})$, and that the square-root function is convex and increasing [53]. Therefore, both $\phi(\mathbf{p})$ and $\varphi(\mathbf{p})$ are convex functions.

E. Proof of Proposition 4

Based on Algorithm 2, we have the inequalities for the zth iteration as follows

$$\begin{aligned} &E_{\text{total}}\left(\boldsymbol{\rho}^{(z-1)}, \mathbf{p}^{(z-1)}, \boldsymbol{\Theta}^{(z-1)}, \boldsymbol{\Phi}^{(z-1)}\right), && (64a) \\ &\ge E_{\text{total}}\left(\boldsymbol{\rho}^{(z)}, \mathbf{p}^{(z-1)}, \boldsymbol{\Theta}^{(z-1)}, \boldsymbol{\Phi}^{(z-1)}\right), && (64b) \\ &\ge E_{\text{total}}\left(\boldsymbol{\rho}^{(z)}, \mathbf{p}^{(z)}, \boldsymbol{\Theta}^{(z-1)}, \boldsymbol{\Phi}^{(z-1)}\right), && (64c) \\ &\ge E_{\text{total}}\left(\boldsymbol{\rho}^{(z)}, \mathbf{p}^{(z)}, \boldsymbol{\Theta}^{(z)}, \boldsymbol{\Phi}^{(z)}\right), && (64d) \end{aligned}$$

where (64b) is due to the convexity of problem (42) and solution $\boldsymbol{\rho}^{(z)}$ represents its optimal solution; (64c) and (64d) are valid according to Proposition 2 and the optimizing phase shifts, respectively. Note that $E_{\text{total}}(\boldsymbol{\rho}, \mathbf{p}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ is decreased at each iteration based on (64a) and (64c). Furthermore, it is obvious that the OF is lower-bounded by a finite value due to constraints. Thus, given a threshold, Algorithm 2 converges within a finite number of iterations.

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, Oct. 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, 3rd Quart., 2017.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[10] T. Bai and R. W. Heath, Jr., "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.

[11] T. Bai, C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for intelligent reflecting surface aided wireless powered mobile edge computing in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5389–5407, Aug. 2021.

[12] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[13] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[14] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[15] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[16] T. Bai, C. Pan, C. Han, and L. Hanzo, "Reconfigurable intelligent surface aided mobile edge computing," *IEEE Wireless Commun.*, 2021. [Online]. Available: https://arxiv.org/pdf/2102.02569.pdf

[17] X. Hu, C. Masouros, and K.-K. Wong, "Removing channel estimation by location-only based deep learning for RIS aided mobile edge computing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

[18] Q. Wu, X. Zhou, and R. Schober, "IRS-assisted wireless powered NOMA: Do we really need different phase shifts in DL and UL?" *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1493–1497, Jul. 2021.

[19] Z. Chu, P. Xiao, M. Shojafar, D. Mi, J. Mao, and W. Hao, "Intelligent reflecting surface assisted mobile edge computing for Internet of Things," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 619–623, Mar. 2021.

[20] K. Wang, Y. Zhou, J. Li, L. Shi, W. Chen, and L. Hanzo, "Energy-efficient task offloading in massive MIMO-aided multi-pair fog-computing networks," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2123–2137, Apr. 2021.

[21] 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.

[22] 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.

[23] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[24] 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.

[25] Z. Liu, Y. Yang, K. Wang, Z. Shao, and J. Zhang, "POST: Parallel offloading of splittable tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3170–3183, Apr. 2020.

[26] 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.

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

[28] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Delay minimization for massive MIMO assisted mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6788–6792, Jun. 2020.

[29] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2884–2899, Apr. 2020.

[30] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Nov. 2020.

[31] Y. Han, W. Tang, S. Jin, C. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, Jun. 2019.

[32] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2020.

[33] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[34] Q. Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, May 2020.

[35] C. Pan *et al.*, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

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

[37] 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.

[38] J. Feng, S. Ma, S. Aissa, and M. Xia, "Two-way massive MIMO relaying systems with non-ideal transceivers: Joint power and hardware scaling," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8273–8289, Dec. 2019.

[39] X. Zhou, B. Bai, and W. Chen, "A low complexity energy efficiency maximization method for multiuser amplify-and-forward MIMO relay systems with a holistic power model," *IEEE Commun. Lett.*, vol. 18, no. 8, pp. 1371–1374, Aug. 2014.

[40] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[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, 4th Quart., 2017.

[42] 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, 3rd Quart., 2017.

[43] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[44] W. Zhao and S. Wang, "Resource allocation for device-to-device communication underlaying cellular networks: An alternating optimization method," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1398–1401, Aug. 2015.

[45] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[46] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[47] A. M.-C. So, J. Zhang, and Y. Ye, "On approximating complex quadratic optimization problems via semidefinite programming relaxations," *Math. Program.*, vol. 110, no. 1, pp. 93–110, 2007.

[48] P. Cao, J. Thompson, and H. V. Poor, "A sequential constraint relaxation algorithm for rank-one constrained problems," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1060–1064.

[49] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[50] D. T. Pham and H. A. L. Thi, "Recent advances in DC programming and DCA," in *Transactions on Computational Intelligence XIII*, vol. 8342. Berlin, Germany: Springer, 2014, pp. 1–37.

[51] H. H. Kha, H. D. Tuan, and H. H. Nguyen, "Fast global optimal power allocation in wireless networks by local D.C. programming," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 510–515, Feb. 2012.

[52] A. Khabbazibasmenj, F. Roemer, S. A. Vorobyov, and M. Haardt, "Sum-rate maximization in two-way AF MIMO relaying: Polynomial time solutions to a class of DC programming problems," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5478–5493, Oct. 2012.

[53] H. Tuy, *Convex Analysis and Global Optimization*. Norwell, MA, USA: Kluwer, 1998.

[54] 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 Symp. Comput. Commun. (ISCC)*, Cappadocia, Turkey, Jul. 2012.

[55] 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.

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. 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.E. 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, Shanghai, China. His research interests include 6G communications, edge intelligence, and the Internet of Things. He was the Track Co-Chair of IEEE VTC 2020 Fall and is the General Co-Chair of IEEE ICC 2022 Workshop on edge artificial intelligence for 6G.

Qingqing Wu (Member, IEEE) received the B.E. degree in electronic engineering from the South China University of Technology and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU) in 2012 and 2016, respectively.

From 2016 to 2020, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore. He is currently an Assistant Professor with the State Key Laboratory of Internet of Things for Smart City, University of Macau. He has coauthored more than 100 IEEE papers with 23 ESI highly cited papers and eight ESI hot papers, which have received more than 9000 Google citations. His current research interests include intelligent reflecting surface (IRS), unmanned aerial vehicle (UAV) communications, and MIMO transceiver design. He was listed among the World's Top 2% Scientist by Stanford University in 2020 and as a Clarivate ESI Highly Cited Researcher in 2021. He was a recipient of the IEEE Communications Society Young Author Best Paper Award in 2021, the Outstanding Ph.D. Thesis Award of China Institute of Communications in 2017, the Outstanding Ph.D. Thesis Funding at SJTU in 2016, the IEEE ICCC Best Paper Award in 2021, and the IEEE WCSP Best Paper Award in 2015. He is the Workshop Co-Chair of IEEE ICC 2019–2022 Workshop on "Integrating UAVs Into 5G and Beyond," and the Workshop Co-Chair of IEEE GLOBECOM 2020 and ICC 2021 Workshop on "Reconfigurable Intelligent Surfaces for Wireless Communication for Beyond 5G." He serves as a Workshops and Symposia Officer of Reconfigurable Intelligent Surfaces Emerging Technology Initiative and a Research Blog Officer of Aerial Communications Emerging Technology Initiative. He is the IEEE Communications Society Young Professional Chair of Asia Pacific Region. He was an Exemplary Editor of IEEE COMMUNICATIONS LETTERS in 2019 and an exemplary reviewer of several IEEE journals. He serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY (OJ-COMS), and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY (OJVT). He is the Lead Guest Editor of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on "UAV Communications in 5G and Beyond Networks," and the Guest Editor of IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY on "6G Intelligent Communications" and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY on "Reconfigurable Intelligent Surface-Based Communications for 6G Wireless Networks."

Wen Chen (Senior Member, IEEE) is a Tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is the Director of the Broadband Access Network Laboratory. He has published more than 110 papers in IEEE journals and more than 100 papers in IEEE conferences, with more than 6000 citations in Google Scholar. His research interests include multiple access, wireless AI, and meta-surface communications. He is a fellow of the Chinese Institute of Electronics and a Distinguished Lecturer of IEEE Communications Society and IEEE Vehicular Technology Society. He is the Shanghai Chapter Chair of IEEE Vehicular Technology Society, an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.

Yang Yang (Fellow, IEEE) received the B.S. and M.S. degrees in radio engineering from Southeast University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong in 2002.

He is currently a Full Professor with the School of Information Science and Technology, a Master at the Kedao College, and the Director of the Shanghai Institute of Fog Computing Technology (SHIFT), ShanghaiTech University, China. He is also an Adjunct Professor with the Research Center for Network Communication, Peng Cheng Laboratory, China, as well as a Senior Consultant at the Shenzhen Smart Cities Technology Development Group, China. Before joining ShanghaiTech University, he has held faculty positions at The Chinese University of Hong Kong, Brunel University London, U.K., University College London (UCL), U.K., and SIMIT, CAS, China. His research interests include 5G/6G, computing networks, service-oriented collaborative intelligence, the IoT applications, and advanced testbeds and experiments. He has published more than 300 papers and filed more than 80 technical patents in these research areas. He has been the Chair of the Steering Committee of Asia-Pacific Conference on Communications (APCC) since January 2019. In addition, he is the General Co-Chair of the IEEE DSP 2018 Conference and the TPC Vice-Chair of the IEEE ICC 2019 Conference.