Joint Communication and Computation Design in Transmissive RMS Transceiver Enabled Multi-Tier Computing Networks

Zhendong Li, Wen Chen, *Senior Member, IEEE*, Ziwei Liu, Hongying Tang, and Jianmin Lu, *Member, IEEE*

***Abstract*— In this paper, a novel transmissive reconfigurable meta-surface (RMS) transceiver enabled multi-tier computing network architecture is proposed for improving computing capability, decreasing computing delay and reducing base station (BS) deployment cost, in which transmissive RMS equipped with a feed antenna can be regarded as a new type of multi-antenna system. We formulate a total energy consumption minimization problem by a joint optimization of subcarrier allocation, task input bits, time slot allocation, transmit power allocation and RMS transmissive coefficient while taking into account the constraints of communication resources and computing resources. This formulated problem is a non-convex optimization problem due to the high coupling of optimization variables, which is NP-hard to obtain its optimal solution. To address the above challenging problems, block coordinate descent (BCD) technique is employed to decouple the optimization variables to solve the problem. Specifically, the joint optimization problem of subcarrier allocation, task input bits, time slot allocation, transmit power allocation and RMS transmissive coefficient is divided into three subproblems to solve by applying BCD. Then, the decoupled three subproblems are optimized alternately by using successive convex approximation (SCA) and difference-convex (DC) programming until the convergence is achieved. Numerical results verify that our proposed algorithm is superior in reducing total energy consumption compared to other benchmarks.**

***Index Terms*— Reconfigurable meta-surface (RMS) transceiver, multi-tier computing network, block coordinate descent (BCD) technique, successive convex approximation (SCA), difference-convex (DC) programming.**

I. Introduction

THE continuous evolution of wireless communications has spawned many emerging applications and services, e.g., telemedicine, industrial Internet and smart Internet-of-Things (IoT) [1], [2]. In these computing and communication-oriented application scenarios, a large number of devices and sensors need to carry out continuous communication and computing, which greatly increases the requirements for devices and sensors. Currently, such computing and communication networks face two main challenges. First, due to the small size of these devices and sensors, the communication, storage and computing capabilities are usually limited, so they cannot support computing-intensive tasks well, which will result in large computing delays and affect users' quality-of-service (QoS) [3]. Therefore, the key issues that the next generation communication network needs to solve are considering how to reduce the computing delay and improve the computing capability of the network. In addition, since the next-generation communication network may use higher frequency bands, the propagation loss will become larger, so its coverage will be reduced. In order to achieve the same coverage as existing communication networks, the number of base stations (BSs) deployed needs to be increased. Moreover, in order to improve the QoS of users, the BS of the next-generation communication network will adopt more antennas, which will increase the required radio frequency (RF) chains, thereby increasing the cost of a single BS. Therefore, how to reduce the deployment cost of the BS is another key challenge that needs to be solved urgently in the next-generation communication network.

A. Related Works

1) MEC Systems: To address the first challenge, powerful computing nodes (CNs) or mobile edge computing (MEC) servers can be deployed at the network edge (i.e. usually co-located with an access point (AP) or BS), which is the recently proposed MEC technology [4], [5], [6], [7], [8]. This technology mainly provides cloud-like computing by deploying MEC servers distributedly in the network, and is widely regarded as an effective means to liberate mobile devices from heavy computing tasks. In the MEC system, devices and sensors with limited computing capability can offload computation-intensive and latency-sensitive tasks to nearby BSs and APs equipped with MEC servers for remote execution, which can greatly reduce computing latency [9]. It is worth noting that the prerequisite for achieving such a goal is that the computing tasks of the devices and sensors can be successfully offloaded. However, since some devices

Manuscript received 11 May 2022; revised 2 September 2022; accepted 25 October 2022. Date of publication 12 December 2022; date of current version 19 January 2023. This work was supported in part by the National Key Project under Grant 2020YFB1807700 and Grant 2018YFB1801102, in part by the Shanghai Kewei under Grant 20JC1416502 and Grant 22JC1404000, in part by Pudong under Grant PKX2021-D02, and in part by NSFC under Grant 62071296. *(Corresponding author: Wen Chen.)*

Zhendong Li, Wen Chen, and Ziwei Liu are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lizhendong@sjtu.edu.cn; wenchen@sjtu.edu.cn; ziweiliu@sjtu.edu.cn).

Hongying Tang is with the Science and Technology on Microsystem Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China (e-mail: tanghy@mail.sim.ac.cn).

Jianmin Lu is with the Wireless Technology Laboratory, Huawei Technologies, Shanghai 201206, China (e-mail: lujianmin@huawei.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3228553.

Digital Object Identifier 10.1109/JSAC.2022.3228553

0733-8716 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

and sensors may be located at the cell edge, their offloading rate is limited, which makes the computing delay at CN or MEC longer than local computing. As a result, these devices and sensors often have to rely on their own resources for computing, which often cannot well support applications for computation-intensive and latency-sensitive tasks. Therefore, it is imperative to improve the offloading capability from the communication perspective, thereby improving the performance of computing and communication networks.

2) Design of MEC Systems: In recent years, the design of MEC communication systems has been widely discussed in academia [10], [11], [12], [13], [14], [15], [16]. Note that in the MEC system, offloading strategies play a crucial role. At present, MEC offloading strategies are mainly divided into binary offloading strategies and partial offloading strategies [17]. The binary offloading strategy mainly decides whether computing tasks are executed locally on devices and sensors or offloaded to CNs or MEC servers for remote execution. The typical tasks used for this offloading strategy are usually simple and indivisible. The partial offloading strategy usually needs to divide the computing task into several sub-tasks, and these sub-tasks can be executed locally through devices and sensors and offloaded to the CNs or MEC servers for parallel execution. Such parallel computing can greatly improve the computing capability and reduce the computing delay of the MEC system. The typical tasks used for this offloading strategy are usually multiple fine-grained processes. In addition, since the rate of offloading will also affect the performance of the MEC system, based on the above two offloading strategies, many research works have studied the joint communication and computing resource allocation in different scenarios to improve the performance of the MEC system.

In previous work, the research on MEC systems can be divided into single-user MEC systems [10], [11], [12], [13] and multi-user MEC systems [4], [14], [15], [16]. In single-user MEC communication systems, Zhang et al. provided a theoretical framework for energy-optimal MEC under stochastic wireless channels by optimizing the execution of mobile applications in mobile devices (i.e., mobile execution) or offloading to the cloud (i.e., cloud execution) to save energy for mobile devices [10]. You et al. proposed an energy-efficient computing framework that includes a set of policies to control CPU cycles for local computing modes, energy transfer, and offload time division for other offloading modes [12]. As for the multi-user MEC system, Ren et al. studied the delay minimization problem in the multi-user time division multiple access system with joint communication and computing resource allocation [14]. Dai et al. proposed a novel two-layer computation offloading framework in heterogeneous networks. Then, in a multi-task MEC system, the joint computation offloading and user association problem is formulated to minimize the overall energy consumption [15]. In addition to the research on the basic MEC system, some emerging communication systems assisted by the MEC are also investigated. Bai et al. studied the application of MEC in unmanned aerial vehicle (UAV) communication networks, and designed an energy-efficient physical layer security optimization algorithm [11]. Liu et al. studied the application of MEC in the Internet-of-Vehicles, and introduced a vehicle fog edge computing paradigm. It is then formulated as a multi-stage Stackelberg game to be solved. However, for multi-user MEC systems, the distribution locations of devices and sensors are usually random and different. Devices and sensors located at the cell edge have a large path loss to the APs or BSs, which will consume more communication resources for offloading, resulting in uneven resource allocation and user fairness issues.

3) RMS Communication Systems: Moreover, faced with the challenge of high cost of BS deployment in next-generation communication networks, reconfigurable meta-surface (RMS) may be a potential solution. RMS has recently been proposed as an emerging technology combining metamaterials and communications, which can be used to reduce network costs, improve network coverage, spectrum- and energy- efficiency [18], [19], [20], [21], [22]. RMS consists of numerous low-cost passive units, and the amplitude and phase shift of the incident signal can be changed by artificially adjusting these units. It is worth noting that since RMS is a passive device, it only adjusts the amplitude and phase of the incident signal, so it will not introduce additional noise, which makes it well applied to a collaborative communication network [19]. In addition, compared with the existing multi-antenna technologies equipped with a large number of RF chains, the hardware cost and power consumption required by passive RMS are much lower, which also greatly stimulates research on RMS-based multi-antenna communication systems. Overall, these advantages mentioned above have greatly promoted the application of RMS in next-generation communication networks [23], [24].

4) Design of RMS Enabled Communication Systems: Based on the above advantages, RMS has attracted extensive attention in academia and industry. Nowadays, RMS is mainly used to assist and enable traditional communication networks, where by combining active-passive beamforming design, the performance of the network can be improved with reflective or transmissive RMS [25], [26], [27], [28], [29], [30], [31], [32], [33]. First, some works on reflective RMS-assisted communication networks has been carried out. Ur Rehman et al. addressed the joint active and passive beamforming optimization problem under ideal and practical IRS phase shifts for an IRS-assisted multi-user downlink MIMO system, where the spectrum efficiency is maximized by minimizing the sum mean squared error (MSE) of the user's received symbols [25]. Zen et al. considered an IRS-assisted uplink non-orthogonal multiple access (NOMA) system in which a semi-definite relaxation technique is employed to maximize the sum rate of users [28]. Furthermore, the research on transmissive RMS-assisted communication networks is also in progress. Zhang et al. proposed an intelligent omni-surface (IOS) assisted downlink communication system, where the IOS is able to forward the received signal to the user in a reflection or transmission manner, thereby enhancing the wireless coverage [30]. Niu et al. investigated a MIMO network assisted by reflection-transmission reconfigurable intelligent surface (RIS), where the weighted sum rate is maximized based on an energy splitting (ES) scheme [33]. Furthermore,

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

in addition to assisting and enabling traditional communication networks, the RMS can also act as transceivers in communication networks. Since the reconfigurability of the RMS helps it to expand the number of passive units without increasing the number of expensive and bulky active antennas, the reflective RMS equipped with an active feed antenna can be used as a new type of transmitter [34]. Since the feed blockage of the transmissive RMS transceiver is less than that of the reflective RMS transceiver, the aperture efficiency can be designed to be higher, and the operating bandwidth can be designed to be more stable, so it is more efficient [35], [36]. At present, some work on the uplink and downlink transmissive RMS transceiver design has been carried out [36], [37], but it is still in its infancy. Meanwhile, the application of the transmissive RMS transceiver in various communication scenarios also has potential.

B. Motivation and Contributions

Based on the above backgrounds and challenges, in order to improve the computing capability, reduce the computing delay, and reduce the BS deployment cost of the communication and computing network, we propose a transmissive RMS transceiver enabled multi-tier computing networks, where the decoding-and-forward (DF) relay is equipped with a CN, and transmissive RMS transceiver is equipped with an MEC server. In general, the computing capability of the DF relay should be comparable to or greater than that of the device to make computational cooperation feasible. To the best of our knowledge, the current research on communication and computing networks with transmissive RMS transceivers is still in its infancy, and the demand for improving network computing capability, reducing computing delay, and reducing BS deployment cost has greatly promoted this work. In this paper, we minimize total energy consumption by jointly optimizing the subcarrier allocation, task input bits, time slot allocation, transmit power allocation, and RMS transmissive coefficient. It is challenging to address this non-convex optimization problem due to the high coupling of optimization variables. Hence, we need to design an effective optimization algorithm for solving it. In summary, the main contributions of this paper can be summarized as follows:

- We propose a novel transmissive RMS transceiver enabled multi-tier computing framework, where the devices and sensors can offload tasks to DF relay and RMS multi-antenna system for calculations, thereby improving computing capability and reducing computing latency of the networks. Meanwhile, we formulate the energy consumption minimization problem for joint optimization of the subcarrier allocation, task input bits, time slot allocation, transmit power allocation, and RMS transmissive coefficient. Since the objective function and the partial constraints are non-convex due to the high coupling of the optimization variables, the problem is a non-convex optimization problem and is challenging to solve directly.
- To address the formulated energy consumption minimization problem, we first divide the non-convex optimization problem into three sub-problems based on the block coordinate descent (BCD) algorithm. Specifically, in the first sub-problem, given the time allocation, task input bits, and RMS transmissive coefficient, we solve the joint optimization problem for the subcarrier allocation and user transmit power allocation. In the second sub-problem, we deal with the joint optimization problem for the time allocation and task input bits by applying successive convex approximation (SCA) when the subcarrier allocation, user transmit power allocation and RMS transmissive coefficient are fixed. For the third sub-problem, the RMS transmissive coefficient can be obtained by using difference-convex (DC) programming and SCA when other optimization variables are given. Finally, the three sub-problems are optimized alternately until convergence is achieved.
- Through the numerical simulation, we verify the effectiveness of the proposed joint optimization algorithm for the subcarrier allocation, task input bits, time slot allocation, transmit power allocation and RMS transmissive coefficient compared with the benchmark algorithms, i.e., it can decrease the total energy consumption. In addition, the proposed multi-layer offload-computation scheme is superior to other offload schemes, and the introduction of transmissive RMS transceivers can greatly reduce the cost of overall network deployment, which has great potential in next-generation communications.

The rest of this paper is organized as follows. Section II elaborates the system model and optimization problem formulation for the transmissive RMS transceiver enabled multi-tier computing networks. Section III presents the proposed optimization algorithm for the formulated optimization problem. In Section IV, numerical results demonstrate that our algorithm has good convergence and effectiveness. Finally, conclusions are given in Section V.

Notations: Scalars are denoted by lower-case letters, while vectors and matrices are represented by bold lower-case letters and bold upper-case letters, respectively. $|x|$ denotes the absolute value of a complex-valued scalar x. For a square matrix $\mathbf{X}$, $\mathrm{tr}(\mathbf{X})$, $\mathrm{rank}(\mathbf{X})$, $\mathbf{X}^H$,$\mathbf{X}_{m,n}$ and $\|\mathbf{X}\|$ denote its trace, rank, conjugate transpose, m,n-th entry and matrix norm, respectively, while $\mathbf{X} \succeq 0$ represents that $\mathbf{X}$ is a positive semidefinite matrix. Similarly, for a general matrix $\mathbf{A}$, $\mathrm{rank}(\mathbf{A})$, $\mathbf{A}^H$, $\mathbf{A}_{m,n}$ and $\|\mathbf{A}\|$ also denote its rank, conjugate transpose, m,n-th entry and matrix norm, respectively. In addition, $\mathbb{C}^{M\times N}$ denotes the space of $M \times N$ complex matrices. $\mathbf{I}_N$ denotes an dentity matrix of size $N \times N$. j denotes the imaginary unit, i.e., $j^2 = -1$. $\mathbb{E}\{\cdot\}$ represents the expectation of random variables. Finally, the distribution of a circularly symmetric complex Gaussian (CSCG) random vector with mean μ and covariance matrix $\mathbf{C}$ is denoted by $\mathcal{CN}(\mu, \mathbf{C})$, and $\sim$ stands for 'distributed as'.

II. System Model and Problem Formulation

In this section, we mainly describe the system model and problem formulation.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

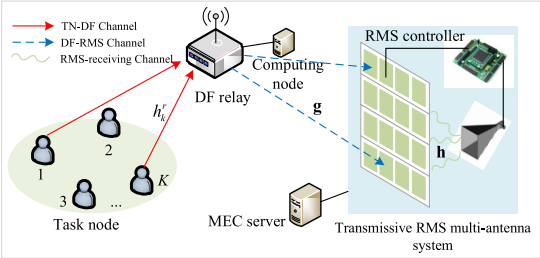


Fig. 1. Transmissive RMS transceiver enabled multi-tier computing networks.

A. Network Model

As shown in the Fig. 1, we consider a multi-tier MEC network model based on a relay-transmissive RMS multi-antenna system, which includes K single-antenna task nodes (TN), a single-antenna DF relay and M transmissive elements RMS multi-antenna system. In this paper, we consider the orthogonal frequency division multiple access (OFDMA) system, where the channel of bandwidth B is divided into N subcarriers, each with a bandwidth of $W = B/N$. Inter-subcarrier interference is negligible, and the cyclic prefix is large enough to overcome inter-symbol interference. Note that TN k needs to successfully execute the $D_k > 0$ task input bits in the time duration $T > 0$. We consider that TN k, DF relay, and RMS multi-antenna systems are all computationally capable. Specifically, the D_k task input bits of TN k can be divided into three parts for computation: local computation, offloading to DF relay computation, and offloading to RMS multi-antenna system computation. Let d_k^l, d_k^r, d_k^m denote the number of task input bits for TN k to be computed locally, offloaded to the DF relay computation, and offloaded to the RMS multi-antenna system computation, respectively. Thus, we have

$$d_k^l + d_k^r + d_k^m = D_k, \quad \forall k. \tag{1}$$

B. Multi-Tier Offloading-Computing Model

We divide the time duration T of TN k for offloading and computing into four-time slots as shown in Fig. 2. In the first time slot $t_k^{\mathrm{I}} \geq 0$, TN k offloads d_k^r task input bits to the DF relay. We assume that the CN and the DF relay are co-located and connected by using high-throughput low-latency optical fibers, so the data transmission between the DF relay and CN can be considered delay-free. The CN executes the task for the remaining $T - t_k^{\mathrm{I}}$ time. In the $t_k^{\mathrm{II}} \geq 0$ and $t_k^{\mathrm{III}} \geq 0$ time slots, TN k offloads the task input bits d_k^m to the RMS multi-antenna system through the DF relay. Specifically, in the second time slot t_k^{II}, TN k sends the task input bits d_k^m to the DF relay. After successfully decoding the received d_k^m, the DF relay forwards it to the RMS multi-antenna system within the third time slot t_k^{III}. The RMS multi-antenna system receives and decodes the signal and sends it to the MEC server. The MEC server executes these tasks in the fourth time slot $t_k^{\mathrm{IV}} \geq 0$. We still assume that the MEC and RMS multi-antenna systems are co-located and connected using high-throughput low-latency fiber, so data transmission between the two can also be considered delay-free.

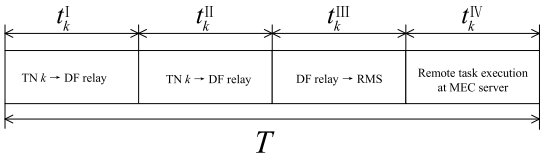


Fig. 2. Multi-tier offloading-computing model.

Since the calculation result bits is usually much smaller than the calculation input bits, the time for the user to download the calculation result is negligible compared to the time for offloading. To ensure that the tasks of TN k can be successfully executed within time duration T, we have the following constraints:

$$t_k^{\mathrm{I}} + t_k^{\mathrm{II}} + t_k^{\mathrm{III}} + t_k^{\mathrm{IV}} \leq T, \quad \forall k. \tag{2}$$

C. Offloading Model

1) Offload to DF Relay: In the time slot t_k^{I}, TN k offloads d_k^r task input bits to the DF relay with transmit power $P_{k,n}^{\mathrm{I}}$. Then the achievable data rate for task offloading of TN k to DF relay on n-th subcarrier can be expressed as

$$r_{k,n}^{\mathrm{I}} = a_{k,n} W \log_2 \left(1 + \frac{P_{k,n}^{\mathrm{I}} \left|h_{k,n}^r\right|^2}{\sigma^2}\right), \quad \forall k, n, \tag{3}$$

where σ^2 represents the power of additive white Gaussian noise (AWGN) introduced at the DF relay. $a_{k,n} \in \{0, 1\}$ indicates whether the n-th subcarrier is allocated to TN k. When the n-th subcarrier is allocated to TN k, $a_{k,n} = 1$. Otherwise, $a_{k,n} = 0$. It should satisfy the following constraints

$$\sum_{k=1}^{K} a_{k,n} \leq 1, \quad \forall n, \tag{4}$$

and $h_{k,n}^r$ represents the channel gain from the TN k to DF relay on n-th subcarrier, which can be modeled as

$$h_{k,n}^r = \sqrt{\frac{C_0}{d_k^\nu}} \left(\sqrt{\frac{\kappa_1}{1+\kappa_1}} e^{-j2\pi n W \frac{d_k}{c}} + \sqrt{\frac{1}{1+\kappa_1}} \tilde{h}_{k,n}\right), \quad \forall k, n \tag{5}$$

with $\tilde{h}_{k,n} \sim \mathcal{CN}(0, 1)$, where C_0 represents the channel gain when the reference distance is 1m, d_k represents the distance from the TN k to the DF relay, ν denotes the path loss coefficient of the corresponding channel, and κ_1 represents the Rician factor of the channel corresponding to TN k. Therefore, in time slot t_k^{I}, the task input bits d_k^r of TN k offloaded to the DF relay can be expressed as

$$d_k^r = t_k^{\mathrm{I}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{I}}, \quad \forall k. \tag{6}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

Let $P_{\max}^{t}$ denote the maximum transmit power of TN k, so we have

$$P_{k,n}^{\mathrm{I}} \geq 0, \quad \forall k, n, \tag{7}$$

and

$$\sum_{n=1}^{N} a_{k,n} P_{k,n}^{\mathrm{I}} \leq P_{\max}^{t}, \quad \forall k. \tag{8}$$

For this offloading process, we regard the transmission energy consumption of TN as the main energy consumption and ignore the energy consumption of circuits such as its radio frequency chain and baseband signal processing. Therefore, in time slot t_k^{I}, the energy consumption of TN k offloading $d_{k,n}^{r}$ task input bits to the DF relay on the n-th subcarrier can be expressed as

$$E_{k,n}^{\mathrm{I}} = a_{k,n} P_{k,n}^{\mathrm{I}} t_k^{\mathrm{I}}, \quad \forall k, n. \tag{9}$$

2) Offlaod to RMS Multi-Antenna System: In the second slot t_k^{II} and the third slot t_k^{III}, the DF relay offloads the task input bits d_k^m of TN k to the RMS multi-antenna system. Let $P_{k,n}^{\mathrm{II}}$ denote the transmit power of TN k on the n-th subcarrier in the time slot t_k^{II}, which should satisfy

$$P_{k,n}^{\mathrm{II}} \geq 0, \quad \forall k, n, \tag{10}$$

and

$$\sum_{n=1}^{N} a_{k,n} P_{k,n}^{\mathrm{II}} \leq P_{\max}^{t}, \quad \forall k. \tag{11}$$

Therefore, the achievable data rate from the TN k to DF relay on the n-th subcarrier at time slot t_k^{II} can be expressed as

$$r_{k,n}^{\mathrm{II}} = a_{k,n} W \log_2 \left(1 + \frac{P_{k,n}^{\mathrm{II}} \left| h_{k,n}^{r} \right|^2}{\sigma^2} \right), \quad \forall k, n. \tag{12}$$

According to the characteristics of the DF relay, after successfully decoding the received signal, it forwards the signal to the RMS multi-antenna system in the third time slot t_k^{III}. Specifically, the RMS multi-antenna system is composed of the receiving antenna and the transmissive RMS. On the n-th subcarrier, the channel from the RMS to the receiving antenna can be expressed as $\mathbf{h}_n \in \mathbb{C}^{M \times 1}$, and the channel from the DF relay to the RMS can be given by $\mathbf{g}_n \in \mathbb{C}^{M \times 1}$. We name them the receiving-RMS channel and the RMS-DF channel, respectively. In this paper, we assume that the channel is constant during the coherence time duration T. Therefore, we only need to obtain three sets of CSI before each optimization, and then apply these CSI in the time duration T. Since the DF relay has the ability to receive and transmit, some classical channel estimation algorithms can be well applied to obtain the CSI from TN k to the DF relay. Then, the DF relay can transmit the obtained CSI to the controller of the RMS transceiver, and the controller performs centralized control. For DF relay to RMS and RMS to receiving antenna, since the receiving antenna has the receiving ability, the transmissive RMS has no receiving ability, the CSI of the cascaded channel can be obtained by drawing on some reflective RMS channel estimation schemes [38], [39]. For the channel from the DF relay to the RMS multi-antenna system, the transmissive RMS adjusts the phase and amplitude of the DF relay forwarded signal and sends it to the receiving antenna. Let $\mathbf{s} = [s_1, \ldots, s_M]^T \in \mathbb{C}^{M \times 1}$ denote the transmissive coefficient vector, where $s_m = \beta_m e^{j\theta_m}, \forall m$. $\beta_m \in [0, 1]$ and $\theta_m \in [0, 2\pi)$ represent the transmissive amplitude and phase shift of the m-th element, respectively. The transmissive coefficient s_m needs to satisfy

$$|s_m| \leq 1, \quad \forall m. \tag{13}$$

For the channel model, the RMS-DF channel is the far-field channel, and the receiving-RMS channel is the near-field channel [37]. We consider the transmissive elements of the RMS to be arranged in uniform planar array (UPA), i.e. $M = M_c \times M_r$, M_c and M_r denote the number of RMS elements on the column and row, respectively. For the RMS-DF channel, we consider that it has both line-of-sight (LoS) and non-line-of-sight (NLoS) components, so we model it as a Rician fading channel, which can be expressed as

$$\mathbf{g}_n = \sqrt{\frac{C_0}{d^{\alpha}}} \left(\sqrt{\frac{\kappa_2}{1+\kappa_2}} e^{-j2\pi n W \frac{\hat{d}}{c}} \mathbf{g}_{\mathrm{LoS}} + \sqrt{\frac{1}{1+\kappa_2}} \mathbf{g}_{\mathrm{NLoS}} \right), \tag{14}$$

with $\mathbf{g}_{\mathrm{NLoS}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M_c M_r})$. $\mathbf{g}_{\mathrm{LoS}}$ can denoted by

$$\begin{aligned} &\mathbf{g}_{\mathrm{LoS}} \\ &= \left[1, e^{-j2\pi f_c \frac{d_r \sin\varphi \cos\psi}{c}}, \ldots, e^{-j2\pi f_c (M_r - 1) \frac{d_r \sin\varphi \cos\psi}{c}} \right]^T \\ &\quad \otimes \left[1, e^{-j2\pi f_c \frac{d_c \sin\varphi \sin\psi}{c}}, \ldots, e^{-j2\pi f_c (M_c - 1) \frac{d_c \sin\varphi \sin\psi}{c}} \right]^T, \end{aligned} \tag{15}$$

where d denotes the distance between the DF relay and RMS, α represents the path loss coefficient of the RMS-DF channel, c represents the speed of light, κ represents the Rician factor and f_c represents the carrier frequency. φ and ψ are the vertical and horizontal angles-of-arrival (AoA) of the incident signal at transmissive RMS, respectively.

Considering that there is no occlusion between the RMS and the receiving antenna, we model the receiving-RMS channel in the near field as a LoS channel, which can be expressed as

$$\mathbf{h}_n = \rho e^{-j2\pi n W \frac{\hat{r}}{c}} \mathbf{h}_{\mathrm{LoS}}, \tag{16}$$

with

$$\mathbf{h}_{\mathrm{LoS}} = \left[e^{-j2\pi f_c \frac{r_{1,1} - \hat{r}}{c}}, \ldots, e^{-j2\pi f_c \frac{r_{M_c, M_r} - \hat{r}}{c}} \right]^H, \tag{17}$$

where ρ and $\hat{r}$ represent the complex channel gain and the distance from the RMS center to the receiving antenna, respectively. The distance from the (m_c, m_r)-th RMS element to the receiving antenna is

$$r_{m_c, m_r} = \sqrt{\hat{r}^2 + \hat{d}_{m_c, m_r}^2}, \tag{18}$$

where $\hat{d}_{m_c, m_r} = \sqrt{\delta_{m_c}^2 d_c^2 + \delta_{m_r}^2 d_r^2}$ denotes the distance from the (m_c, m_r)-th element of the RMS to the RMS center. d_c and d_r denote the column spacing and row spacing from the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

(m_c, m_r)-th element to the center element, respectively. $\delta_{m_c} = \frac{2m_c - M_c - 1}{2}$ and $\delta_{m_r} = \frac{2m_r - M_r - 1}{2}$.

Therefore, in the time slot t_k^{III}, the achievable rate of the signal forwarded by the DF relay received by the receiving antenna is

$$r_{k,n}^{\mathrm{III}} = b_{k,n} W \log_2 \left(1 + \frac{P_{k,n}^{\mathrm{III}} \left|\mathbf{h}_n^H \mathrm{diag}\left(\mathbf{g}_n\right)\mathbf{s}\right|^2}{\delta^2}\right), \quad \forall k, n, \tag{19}$$

where δ^2 represents the power of AWGN introduced at the feeding antenna, $b_{k,n} \in \{0, 1\}$ denotes the subcarrier allocation variable, which is constrained by the following

$$\sum_{k=1}^{K} b_{k,n} \leq 1, \quad \forall n, \tag{20}$$

and $P_{k,n}^{\mathrm{III}}$ represents the TN k transmit power allocated on the n-th subcarrier by the DF relay in the time slot t_k^{III}, which should satisfy

$$P_{k,n}^{\mathrm{III}} \geq 0, \quad \forall k, n, \tag{21}$$

and

$$\sum_{k=1}^{K} \sum_{n=1}^{N} b_{k,n} P_{k,n}^{\mathrm{III}} \leq P_{\max}^{r}, \tag{22}$$

where $P_{\max}^{r}$ denotes the maximum transmit power of DF relay.

According to the achievable rate of TN k in the time slot t_k^{II} and t_k^{III}, the task input bits d_k^m of TN k offloaded to RMS multi-antenna system through DF relay needs to satisfy

$$d_k^m = \min\left(t_k^{\mathrm{II}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{II}}, t_k^{\mathrm{III}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{III}}\right), \quad \forall k. \tag{23}$$

We consider the transmission energy consumption of TN k and DF relay for offloading as the main energy consumption in the time slot t_k^{II} and t_k^{III}. We have

$$E_{k,n}^{\mathrm{II}} = a_{k,n} P_{k,n}^{\mathrm{II}} t_k^{\mathrm{II}}, \quad \forall k, n, \tag{24}$$

and

$$E_{k,n}^{\mathrm{III}} = b_{k,n} P_{k,n}^{\mathrm{III}} t_k^{\mathrm{III}}, \quad \forall k, n. \tag{25}$$

Therefore, the offloading energy consumption of TN k offloaded to the DF relay and to the RMS multi-antenna system through the DF relay can be expressed as

$$E_k^{off} = \sum_{n=1}^{N} \left(E_{k,n}^{\mathrm{I}} + E_{k,n}^{\mathrm{II}} + E_{k,n}^{\mathrm{III}}\right), \quad \forall k. \tag{26}$$

D. Computing Model

1) TN Local Computing Model: During the time duration T, TN k executes d_k^l task input bits. In fact, the number of CPU cycles to perform a computing task depends largely on various factors, e.g., the specific application, the number of task input bits, and the hardware device used for the computation. To characterize the most necessary computation and communication tradeoff, we consider the number of CPU cycles to execute a task as a linear function of the number of task input bits, where c_t represents the number of CPU cycles to compute each task input bit at TN. Let $f_{t,i}$ denote the CPU frequency of the i-th cycle of TN, which is subject to the following constraints:

$$f_{t,i} \leq f_{t,\max}, \quad \forall i \in \left\{1, \ldots, c_t d_k^l\right\}, \tag{27}$$

where $f_{t,\max}$ represents the maximum CPU frequency when the TN executes the task. Since the task input bits computed locally by TN k should be successfully executed within time duration T, we have the following delay constraints:

$$\sum_{i=1}^{c_t d_k^l} \frac{1}{f_{t,i}} \leq T. \tag{28}$$

Therefore, the local computing energy consumption of TN k can be expressed as

$$E_k^{t,comp} = \sum_{i=1}^{c_t d_k^l} \alpha_t f_{t,i}^2, \tag{29}$$

where α_t depends on the effective capacitance factor of TN chip architecture. In order to save computing power consumption under computating latency constraints, it is better to set the CPU frequency to be the same for different CPU cycles [40]. By using this fact and making the constraints in Eq. (28) satisfy strict equality (in order to minimize computing energy consumption), we have

$$f_{t,1} = f_{t,2} = \ldots = f_{t,c_t d_k^l} = \frac{c_t d_k^l}{T}, \quad \forall k. \tag{30}$$

Therefore, the local computing energy consumption of TN k can be further expressed as

$$E_k^{t,comp} = \frac{\alpha_t \left(c_t d_k^l\right)^3}{T^2}, \quad \forall k. \tag{31}$$

Additionally, the CPU frequency for local computation is bound by the maximum CPU frequency that can be given by

$$\frac{c_t d_k^l}{T} \leq f_{t,\max}, \quad \forall k. \tag{32}$$

2) CN Computing Model: After receiving the d_k^r task input bits offloaded by TN k in the first time slot t_k^{I}, the DF relay executes the calculation through the connected CN. We assume that the two are co-located and connected using high-throughput, low-latency fiber, so their transmission delays are negligible. CN executes tasks within the remaining $T - t_k^{\mathrm{I}}$. Let $f_{r,i}$ and $f_{r,\max}$ denote the CPU frequency and the maximum frequency of the CPU in the i-th cycle of CN, respectively. Similar to the local computing of TN k, we have

$$f_{r,i} = \frac{c_r d_k^r}{T - t_k^{\mathrm{I}}}, \quad \forall i \in \left\{1, \ldots, c_r d_k^r\right\}, \tag{33}$$

where c_r represents the number of CPU cycles that CN executes each task input bit. Therefore, the computing energy consumption of CN to execute d_k^r task bits offloaded by TN k can be expressed as

$$E_k^{r,comp} = \frac{\alpha_r (c_r d_k^r)^3}{\left(T - t_k^{\mathrm{I}}\right)^2}, \quad \forall k, \tag{34}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

where α_r depends on the effective capacitance factor of CN chip architecture. In order to ensure that the CN can successfully execute the task, the task input bits should satisfy the following constraints

$$\sum_{k=1}^{K} \frac{c_r d_k^r}{T - t_k^{\mathrm{I}}} \leq f_{r,\max}, \tag{35}$$

where $f_{r,\max}$ represents the maximum CPU frequency when CN executes tasks.

3) MEC Computing Model: In the fourth time slot t_k^{IV}, the RMS multi-antenna system receives the d_k^m task input bits offloaded of TN k by the DF relay on the n-th subcarrier, and then forwards it to the nearby MEC server for computing. Similarly, we assume that the two are co-located and connected using high-throughput, low-latency fiber, so their transmission delays are also negligible. Therefore, the time required for the MEC server to execute d_k^m task input bits is expressed as

$$t_k^{\mathrm{IV}} = \frac{c_m d_k^m}{f_{m,i}}, \quad \forall i \in \{1, \ldots, c_r d_k^m\}, \tag{36}$$

where c_m represents the number of CPU cycles that MEC executes each task input bit. Similarly, in order to ensure that the MEC can successfully execute the task, there are the following constraints on the task input bits

$$\sum_{k=1}^{K} \frac{c_m d_k^m}{t_k^{\mathrm{IV}}} \leq f_{m,\max}. \tag{37}$$

E. Problem Formulation

Considering that TN and DF relay are usually wireless devices, the energy management of them is more complicated, and the RMS transceiver as a BS usually has a reliable power supply, so we temporarily consider the energy consumption of wireless device side (i.e., TN and DF relay) as the main performance indicator. In this paper, we aim to minimize the total energy consumption of TN and DF relay in the transmissive RMS transceiver enabled multi-tier computing networks by jointly optimizing the subcarrier allocation **A (B)**, task input bits **D**, time slot allocation **T**, transmit power allocation **P** and RMS transmissive coefficient **s**. This optimization problem can be formulated as

$$\text{(P0)} \quad \min_{\mathbf{A},\mathbf{B},\mathbf{D},\mathbf{T},\mathbf{P},\mathbf{s}} \sum_{k=1}^{K} \left(E_k^{off} + E_k^{t,comp} + E_k^{r,comp} \right), \tag{38a}$$

$$s.t.\ d_k^r \leq t_k^{\mathrm{I}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{I}}, \quad \forall k, \tag{38b}$$

$$d_k^m \leq \min \left(t_k^{\mathrm{II}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{II}}, t_k^{\mathrm{III}} \sum_{n=1}^{N} r_{k,n}^{\mathrm{III}} \right), \ \forall k, \tag{38c}$$

$$d_k^l + d_k^r + d_k^m = D_k, \quad \forall k, \tag{38d}$$

$$P_{k,n}^S \geq 0, S \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}, \quad \forall k, n, \tag{38e}$$

$$\sum_{n=1}^{N} a_{k,n} P_{k,n}^S \leq P_{\max}^t, S \in \{\mathrm{I}, \mathrm{II}\}, \quad \forall k, \tag{38f}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} b_{k,n} P_{k,n}^{\mathrm{III}} \leq P_{\max}^r, \tag{38g}$$

$$a_{k,n}, b_{k,n} \in \{0, 1\}, \quad \forall k, n, \tag{38h}$$

$$\sum_{k=1}^{K} a_{k,n} \leq 1, \quad \forall n, \tag{38i}$$

$$\sum_{k=1}^{K} b_{k,n} \leq 1, \quad \forall n, \tag{38j}$$

$$\frac{c_t d_k^l}{T} \leq f_{t,\max}, \quad \forall k, \tag{38k}$$

$$\sum_{k=1}^{K} \frac{c_r d_k^r}{T - t_k^{\mathrm{I}}} \leq f_{r,\max}, \tag{38l}$$

$$\sum_{k=1}^{K} \frac{c_m d_k^m}{t_k^{\mathrm{IV}}} \leq f_{m,\max}, \tag{38m}$$

$$t_k^S \geq 0, S \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}\}, \quad \forall k, \tag{38n}$$

$$t_k^{\mathrm{I}} + t_k^{\mathrm{II}} + t_k^{\mathrm{III}} + t_k^{\mathrm{IV}} \leq T, \quad \forall k, \tag{38o}$$

$$|s_m| \leq 1, \quad \forall m, \tag{38p}$$

where (38b)-(38d) represent the task input bit constraints for TN k offload to DF relay and RMS multi-antenna system. (38e)-(38g) denote the transmit power constraints for TN k and DF relay. (38h)-(38j) represent subcarrier allocation constraints in different time slots. (38k)-(38m) denote the computing capability constraints of TN k, CN, and MEC server. (38n) and (38o) represent time allocation constraints. (38p) represents the transmissive coefficient constraint of the RMS multi-antenna system. Note that this joint optimization problem is designed offline, i.e., we obtain the optimization variable $\{\mathbf{A}, \mathbf{B}, \mathbf{D}, \mathbf{T}, \mathbf{P}, \mathbf{s}\}$ under the assumption that the user location and all channel state information (CSI) are all perfectly acquired.

However, the solution to the problem (P0) is challenging for the following reasons. Firstly, the optimization variables are highly coupled, which leads to the objective function and constraints being non-convex with respect to (w.r.t) the optimization variables. Then, the constraints (38f)-(38j) involve binary variables. Therefore, the problem (P0) is a mixed-integer non-convex optimization problem, and it is quite challenging to obtain the global optimal solution. Consequently, we need to design an efficient algorithm to obtain a high-quality sub-optimal solution through the BCD algorithm. The algorithm for solving the problem (P0) will be introduced in detail below.

III. Joint Optimization Algorithm for the Transmissive RMS Transceiver Enabled Multi-Tier Computing Networks

In this section, since the formulated problem (P0) is a non-convex optimization problem, we divide the problem (P0) into three sub-problems for solving based on the BCD algorithm. Specifically, In the sub-problem 1, the time slot allocation **T**, the task input bit **D** and the RMS transmissive coefficient **s** are fixed, and the subcarrier allocation **A**, **B** and the transmit power allocation **P** are jointly optimized. In the sub-problem 2, given the subcarrier allocation **A**, **B**, the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

transmit power **P** and the RMS transmissive coefficient **s**, the task input bit **D** and time allocation **T** are jointly optimized. The third sub-problem is to optimize the RMS transmissive coefficient **s** for fixed subcarrier allocation **A**, **B**, transmit power allocation **P**, task input bit **D** and time allocation **T**. Finally, the three sub-problems are optimized alternately until convergence is achieved.

The existence of binary variables **A** and **B** makes constraints (38f)-(38j) non-convex constraints. In order to solve this problem, we first relax the binary variable $a_{k,n}$ to obtain $\tilde{a}_{k,n}$, i.e., $a_{k,n} \in \{0,1\} \Rightarrow \tilde{a}_{k,n} \in [0,1], \forall k,n$. Then, the auxiliary variables $\tilde{P}^{\mathrm{I}}_{k,n} = \tilde{a}_{k,n} P^{\mathrm{I}}_{k,n}, \forall k,n$ and $\tilde{P}^{\mathrm{II}}_{k,n} = \tilde{a}_{k,n} P^{\mathrm{II}}_{k,n}, \forall k,n$ are introduced. Similarly, for the binary variable $b_{k,n}$, we have $b_{k,n} \in \{0,1\} \Rightarrow \tilde{b}_{k,n} \in [0,1], \forall k,n$, and then introduce the auxiliary variable $\tilde{P}^{\mathrm{III}}_{k,n} = \tilde{b}_{k,n} P^{\mathrm{III}}_{k,n}, \forall k,n$. After variable relaxation and the introduction of auxiliary variables, the optimization problem (P0) can be expressed as the optimization problem (P1) as follows

$$
\begin{aligned}
(\mathrm{P1}) \quad & \min_{\tilde{\mathbf{A}},\tilde{\mathbf{B}},\mathbf{D},\mathbf{T},\tilde{\mathbf{P}},\mathbf{s}} \sum_{k=1}^{K} \\
& \left(\sum_{n=1}^{N} \left(\tilde{P}^{\mathrm{I}}_{k,n} t^{\mathrm{I}}_k + \tilde{P}^{\mathrm{II}}_{k,n} t^{\mathrm{II}}_k + \tilde{P}^{\mathrm{III}}_{k,n} t^{\mathrm{III}}_k \right) \right. \\
& \left. + E^{t,comp}_k + E^{r,comp}_k \right), && (39a) \\
& s.t. \ \text{(38b)-(38d), (38k)-(38p)}, && (39b) \\
& \tilde{P}^{S}_{k,n} \geq 0, S \in \{\mathrm{I},\mathrm{II},\mathrm{III}\}, \ \forall k,n, && (39c) \\
& \sum_{n=1}^{N} \tilde{P}^{S}_{k,n} \leq P^{t}_{\max}, S \in \{\mathrm{I},\mathrm{II}\}, \quad \forall k, && (39d) \\
& \sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{P}^{\mathrm{III}}_{k,n} \leq P^{r}_{\max}, && (39e) \\
& \tilde{a}_{k,n}, \tilde{b}_{k,n} \in [0,1], \quad \forall k,n, && (39f) \\
& \sum_{k=1}^{K} \tilde{a}_{k,n} \leq 0, \quad \forall n, && (39g) \\
& \sum_{k=1}^{K} \tilde{b}_{k,n} \leq 0, \quad \forall n. && (39h)
\end{aligned}
$$

A. Sub-Problem 1: Joint Optimization of Subcarrier Allocation and Transmit Power Allocation

In this subsection, firstly, given the time allocation **T**, the task input bit **D** and the RMS transmissive coefficient **s**, the subcarrier allocation $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ and transmit power allocation $\tilde{\mathbf{P}}$ are jointly optimized, then the problem (P1) can be written as the problem (P2), which is expressed as

$$
\begin{aligned}
(\mathrm{P2}) \quad & \min_{\tilde{\mathbf{A}},\tilde{\mathbf{B}},\tilde{\mathbf{P}}} \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \left(\tilde{P}^{\mathrm{I}}_{k,n} t^{\mathrm{I}}_k + \tilde{P}^{\mathrm{II}}_{k,n} t^{\mathrm{II}}_k + \tilde{P}^{\mathrm{III}}_{k,n} t^{\mathrm{III}}_k \right) \right. \\
& \left. + \frac{\alpha_t \left(c_t d^l_k \right)^3}{T^2} + \frac{\alpha_t (c_r d^r_k)^3}{\left(T - t^{\mathrm{I}}_k \right)^2} \right), && (40a) \\
& s.t. \ \text{(38b), (39c)-(39h)}, && (40b) \\
& d^m_k \leq t^{\mathrm{II}}_k \sum_{n=1}^{N} r^{\mathrm{II}}_{k,n}, \quad \forall k, && (40c) \\
& d^m_k \leq t^{\mathrm{III}}_k \sum_{n=1}^{N} r^{\mathrm{III}}_{k,n}, \quad \forall k, && (40d)
\end{aligned}
$$

Lemma 1: The function $f(x,t) = t\log_2\left(1 + \frac{x}{t}\right)$ is concave w.r.t $x > 0$ and $t > 0$.

Proof: The function $f(x,t)$ can be obtained by the perspective transformation of the function $h(x) = \log_2(1+x)$, i.e, $f(x,t) = th\left(\frac{x}{t}\right)$. Since the perspective function is concave-preserving and $h(x) = \log_2(1+x)$ is concave function w.r.t $x > 0$, the function $f(x,t) = t\log_2\left(1 + \frac{x}{t}\right)$ is concave w.r.t $x > 0$ and $t > 0$. The proof of **Lemma 1** is completed. ■

According to the **Lemma 1**, $r^{\mathrm{I}}_{k,n}$ is concave w.r.t $\tilde{a}_{k,n}$ and $\tilde{P}^{\mathrm{I}}_{k,n}$, $r^{\mathrm{II}}_{k,n}$ is concave w.r.t $\tilde{a}_{k,n}$ and $\tilde{P}^{\mathrm{II}}_{k,n}$, and $r^{\mathrm{III}}_{k,n}$ is concave w.r.t $\tilde{b}_{k,n}$ and $\tilde{P}^{\mathrm{III}}_{k,n}$. Hence, (38b), (40c) and (40d) are convex constraints. In addtion, the objective function is affine w.r.t optimization variables and (39c)-(39h) are also affine constraints. Thus, it can be seen that the problem (P2) is a standard convex optimization problem, which can be solved by using CVX toolbox [41].

Remark: To facilitate obtaining the solution to the problem (P2), the subcarrier allocation variables $a_{k,n}$ and $b_{k,n}$ are relaxed into continuous variables. After obtaining the subcarrier allocation variable value, it needs to be restored to obtain the subcarrier allocation scheme. Considering that one user can occupy multiple subcarriers, but one subcarrier can only be allocated to one user, we obtain the maximum value of all user subcarrier allocation variables on the n-th subcarrier and set the $a_{k,n}$ or $b_{k,n}$ of the corresponding user to 1, and the $a_{k,n}$ or $b_{k,n}$ of the remaining users to 0.

B. Sub-Problem 2: Joint Optimization of Task Input Bit and Time Allocation

In this subsection, given subcarrier allocation **A**, **B**, transmit power allocation **P** and RMS transmissive coefficient **s**, we jointly optimize task input bit **D** and time allocation **T**, then the problem (P1) can be transformed into the problem (P3), which can be expressed as

$$
\begin{aligned}
(\mathrm{P3}) \quad & \min_{\mathbf{D},\mathbf{T}} \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \left(\tilde{P}^{\mathrm{I}}_{k,n} t^{\mathrm{I}}_k + \tilde{P}^{\mathrm{II}}_{k,n} t^{\mathrm{II}}_k + \tilde{P}^{\mathrm{III}}_{k,n} t^{\mathrm{III}}_k \right) \right. \\
& \left. + \frac{\alpha_t \left(c_t d^l_k \right)^3}{T^2} + \frac{\alpha_t (c_r d^r_k)^3}{\left(T - t^{\mathrm{I}}_k \right)^2} \right), && (41a) \\
& s.t. \ \text{(38b), (38d), (38k)-(38o), (40c), (40d).} && (41b)
\end{aligned}
$$

Lemma 2: $E^{r,comp}_k = \frac{\alpha_t (c_r d^r_k)^3}{\left(T - t^{\mathrm{I}}_k \right)^2}$ is a convex function w.r.t $d^r_k > 0$ and $t^{\mathrm{I}}_k > 0$.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

Proof: The Hessian matrix of $E_k^{r,comp}=\frac{\alpha_t(c_r d_k^r)^3}{\left(T-t_k^{\mathrm{I}}\right)^2}$ is

$$\begin{bmatrix} \frac{6{c_r}^3\alpha_t d_k^r}{\left(T-t_k^{\mathrm{I}}\right)^2} & \frac{6{c_r}^3\alpha_t (d_k^r)^2}{\left(T-t_k^{\mathrm{I}}\right)^3} \\ \frac{6{c_r}^3\alpha_t (d_k^r)^2}{\left(T-t_k^{\mathrm{I}}\right)^3} & \frac{6{c_r}^3\alpha_t (d_k^r)^3}{\left(T-t_k^{\mathrm{I}}\right)^4} \end{bmatrix}. \tag{42}$$

Its eigenvalues are 0 and $\frac{6{c_r}^3\alpha_t d_k^r}{\left(T-t_k^{\mathrm{I}}\right)^2}\left(1+\frac{(d_k^r)^2}{\left(T-t_k^{\mathrm{I}}\right)^2}\right)$, so the matrix is a semi-positive definite matrix, so $E_k^{r,comp}$ is jointly convex w.r.t $d_k^r>0$ and $t_k^{\mathrm{I}}>0$. This completes the proof of **Lemma 2**. ■

In the objective function of problem (P3), it can be proved that E_k^{off} is affine w.r.t optimization variables, and $E_k^{t,comp}$ is convex w.r.t optimization variables. According to **Lemma 2**, $E_k^{r,comp}$ is also a convex function w.r.t the optimization variables. Therefore, the objective function of this problem (P3) is convex. In addition, (38b), (38d), (38k), (38n), (38o), (40c) and (40d) are affine. (38l) and (38m) are non-convex constraints, which makes the problem (P3) still a non-convex optimization problem. Next, we can apply SCA to carry out the first-order Taylor expansion of the left-hand-side (LHS) concave functions of constraints (38l) and (38m), and obtain their upper bounds respectively.[1]

Specifically, for the LHS $\frac{c_r d_k^r}{T-t_k^{\mathrm{I}}}$ of constraint (38l), we adopt SCA to obtain its upper bound, which can be expressed as

$$\begin{aligned}\frac{c_r d_k^r}{T-t_k^{\mathrm{I}}} \le{}& \frac{c_r d_k^{r(i)}}{T-t_k^{\mathrm{I}(i)}}+\frac{c_r}{T-t_k^{\mathrm{I}(i)}}\left(d_k^r-d_k^{r(i)}\right)\\ &+\frac{c_r d_k^{r(i)}}{\left(T-t_k^{\mathrm{I}(i)}\right)^2}\left(t_k^{\mathrm{I}}-t_k^{\mathrm{I}(i)}\right)\triangleq\left(\frac{c_r d_k^r}{T-t_k^{\mathrm{I}}}\right)^{ub},\end{aligned} \tag{43}$$

where $d_k^{r(i)}$ and $t_k^{\mathrm{I}(i)}$ represent the values of d_k^r and t_k^{I} at the i-th SCA iteration, respectively. Similarly, for the LHS $\frac{c_m d_{k,n}^m}{t_k^{\mathrm{IV}}}$ of constraint (38m), we also adopt SCA to obtain its upper bound, which can be expressed as

$$\begin{aligned}\frac{c_m d_{k,n}^m}{t_k^{\mathrm{IV}}} \le{}& \frac{c_m d_{k,n}^{m(i)}}{t_k^{\mathrm{IV}(i)}}+\frac{c_m}{t_k^{\mathrm{IV}(i)}}\left(d_{k,n}^m-d_{k,n}^{m(i)}\right)\\ &-\frac{c_m d_{k,n}^{m(i)}}{\left(t_k^{\mathrm{IV}(i)}\right)^2}\left(t_k^{\mathrm{IV}}-t_k^{\mathrm{IV}(i)}\right)\triangleq\left(\frac{c_m d_{k,n}^m}{t_k^{\mathrm{IV}}}\right)^{ub},\end{aligned} \tag{44}$$

where $d_{k,n}^{m(i)}$ and $t_k^{\mathrm{IV}(i)}$ denote the values of $d_{k,n}^m$ and t_k^{IV} at the i-th SCA iteration, respectively. Therefore, the problem (P3) can be approximately transformed into the problem (P3.1), which can be expressed as

$$\begin{aligned}(\mathrm{P3.1})\quad \min_{\mathbf{D},\mathbf{T}}\ & \sum_{k=1}^{K}\left(\sum_{n=1}^{N}\left(\tilde{P}_{k,n}^{\mathrm{I}}t_k^{\mathrm{I}}+\tilde{P}_{k,n}^{\mathrm{II}}t_k^{\mathrm{II}}+\tilde{P}_{k,n}^{\mathrm{III}}t_k^{\mathrm{III}}\right)+\right.\\ &\left.\frac{\alpha_t\left(c_t d_k^l\right)^3}{T^2}+\frac{\alpha_t(c_r d_k^r)^3}{\left(T-t_k^{\mathrm{I}}\right)^2}\right),\end{aligned} \tag{45a}$$

$$s.t.\ \text{(38b), (38d), (38k), (38n), (38o), (40c), (40d),} \tag{45b}$$

$$\sum_{k=1}^{K}\left(\frac{c_r d_k^r}{T-t_k^{\mathrm{I}}}\right)^{ub}\le f_{r,\max}, \tag{45c}$$

$$\sum_{k=1}^{K}\left(\frac{c_m d_k^m}{t_k^{\mathrm{IV}}}\right)^{ub}\le f_{m,\max}. \tag{45d}$$

The problem (P3.1) is a standard convex optimization problem that can be solved by using the CVX toolbox [41].

C. Sub-Problem 3: Optimization of RMS Transmissive Coefficient

In this subsection, we fix the subcarrier allocation **A**, **B**, the transmit power allocation **P**, the task input bit **D** and the time allocation **T**, and solve the transmissive coefficient of RMS. Since the objective function does not contain the RMS transmissive coefficient vector explicitly, the problem (P1) can be transformed into feasibility-check problem (P4), which can be expressed as

$$(\mathrm{P4})\quad \text{find}\ \ \mathbf{s}, \tag{46a}$$

$$s.t.\ d_k^m\le t_k^{\mathrm{III}}\sum_{n=1}^{N}r_{k,n}^{\mathrm{III}},\quad \forall k, \tag{46b}$$

$$|s_m|\le 1,\quad \forall m. \tag{46c}$$

It can be seen that the problem (P4) is a non-convex optimization problem. To solve this problem, we let $\mathbf{v}_n^H=\mathbf{h}_n^H\mathrm{diag}\left(\mathbf{g}_n\right)\in\mathbb{C}^{1\times M}$, then $\left|\mathbf{h}_n^H\mathrm{diag}\left(\mathbf{g}_n\right)\mathbf{s}\right|^2=\left|\mathbf{v}_n^H\mathbf{s}\right|^2=\mathbf{v}_n^H\mathbf{s}\mathbf{s}^H\mathbf{v}_n$. Let $\mathbf{S}=\mathbf{s}\mathbf{s}^H\in\mathbb{C}^{M\times M}$, $\mathbf{S}\succeq 0$ and $\mathrm{rank}\left(\mathbf{S}\right)=1$. In addition, we let $\mathbf{V}_n=\mathbf{v}_n\mathbf{v}_n^H\in\mathbb{C}^{M\times M}$, then $\left|\mathbf{h}_n^H\mathrm{diag}\left(\mathbf{g}_n\right)\mathbf{s}\right|^2=\mathrm{tr}\left(\mathbf{S}\mathbf{V}_n\right)$. Therefore, $r_{k,n}^{\mathrm{III}}$ can be equivalently expressed as

$$r_{k,n}^{\mathrm{III}}=b_{k,n}W\log_2\left(1+\frac{P_{k,n}^{\mathrm{III}}\mathrm{tr}\left(\mathbf{S}\mathbf{V}_n\right)}{\delta^2}\right),\quad \forall k,n. \tag{47}$$

Then the problem (P4) can be equivalently expressed as the problem (P4.1), which can be given by

$$(\mathrm{P4.1})\quad \text{find}\ \ \mathbf{S}, \tag{48a}$$

$$s.t.\ \text{(46b)}, \tag{48b}$$

$$\mathbf{S}_{m,m}\le 1,\quad \forall m, \tag{48c}$$

$$\mathbf{S}\succeq 0, \tag{48d}$$

$$\mathrm{rank}\left(\mathbf{S}\right)=1. \tag{48e}$$

It can be seen that the problem (P4.1) is a non-convex optimization problem due to the existence of the non-convex rank-one constraint (46e). Next, we apply **Proposition 1** to transform constraint (46e).

Proposition 1: For any positive semi-definite matrix $\mathbf{C}\in\mathbb{C}^{N\times N}$, $\mathrm{tr}\left(\mathbf{C}\right)>0$, the rank-one constraint can be equivalently expressed as

$$\mathrm{rank}\left(\mathbf{C}\right)=1\Rightarrow\mathrm{tr}\left(\mathbf{C}\right)-\left\|\mathbf{C}\right\|_2=0, \tag{49}$$

where $\mathrm{tr}\left(\mathbf{C}\right)=\sum_{n=1}^{N}\sigma_n\left(\mathbf{C}\right)$, $\left\|\mathbf{C}\right\|_2=\sigma_1\left(\mathbf{C}\right)$ represents the spectral norm of the matrix $\mathbf{C}$. $\sigma_n\left(\mathbf{C}\right)$ denotes the n-th largest singular value of matrix $\mathbf{C}$.

[1]According to the second-order condition of convex functions, the LHS of Eq. (38l) and Eq. (38m) are concave functions, which are easy to prove and are omitted here.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

According to **Proposition 1**, we can transform the rank-one constraint (48e) in the problem (P4.1) into

$$\mathrm{rank}\,(\mathbf{S}) = 1 \Rightarrow \mathrm{tr}\,(\mathbf{S}) - \|\mathbf{S}\|_2 = 0. \tag{50}$$

Thus, the feasible-check problem (P4.1) can be transformed into the problem (P4.2), which can be expressed as

$$(\mathrm{P4.2}) \qquad \min_{\mathbf{S}} \ \mathrm{tr}\,(\mathbf{S}) - \|\mathbf{S}\|_2, \tag{51a}$$

$$s.t.\ \text{(46b), (48c), (48d).} \tag{51b}$$

Since $\|\mathbf{S}\|_2$ is a convex function, the problem (P4.2) is still a non-convex optimization problem. Here, we use SCA to obtain the lower bound of $\|\mathbf{S}\|_2$, which can be expressed as

$$\begin{aligned}\|\mathbf{S}\|_2 &\geq \left\|\mathbf{S}^{(i)}\right\|_2 + \mathrm{tr}\left(\mathbf{u}_{\max}\left(\mathbf{S}^{(i)}\right)\mathbf{u}_{\max}\left(\mathbf{S}^{(i)}\right)^H\left(\mathbf{S}-\mathbf{S}^{(i)}\right)\right)\\ &\triangleq \left(\|\mathbf{S}\|_2\right)^{lb},\end{aligned} \tag{52}$$

where $\mathbf{u}_{\max}\left(\mathbf{S}^{(i)}\right)$ denotes the eigenvector corresponding to the largest singular value of the matrix $\mathbf{S}^{(i)}$, and $\mathbf{S}^{(i)}$ represents the value of $\mathbf{S}$ in the i-th SCA iteration. Therefore, the problem (P4.2) can be approximately converted to the problem (P4.3), which can be given by

$$(\mathrm{P4.3}) \qquad \min_{\mathbf{S}} \ \mathrm{tr}\,(\mathbf{S}) - \left(\|\mathbf{S}\|_2\right)^{lb}, \tag{53a}$$

$$s.t.\ \text{(46b), (48c), (48d).} \tag{53b}$$

It can be seen that the problem (P4.3) is a standard SDP problem, which can be solved by using CVX toolbox [41].

D. The Overall Joint Optimization Algorithm in Multi-Tier Computing Networks

Since multiple optimization variables are highly coupled, the original problem is a complex non-convex optimization problem. In this paper, we decouple the original problem into three sub-problems to solve through the BCD algorithm framework. Specifically, in sub-problem 1, we first relax the binary variables, and then given the task input bits **D**, the time allocation **T**, and the RMS transmissive coefficient **s**, we can obtain the subcarrier allocation $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$, and the transmit power allocation $\tilde{\mathbf{P}}$. Finally, the binary variable is restored. In sub-problem 2, given subcarrier allocation **A**, **B**, transmit power allocation **P**, and RMS transmissive coefficient **s**, we obtain task input bits **D** and time allocation **T** by applying SCA technique. In sub-problem 3, with fixed subcarrier allocation **A**, **B**, transmit power allocation **P**, task input bits **D**, and time allocation **T**, we can obtain RMS transmissive coefficient **s** by applying SCA and DC programming. Finally, the three subproblems are optimized alternately until the convergence is achieved. Based on the solution of the above subproblems, we propose a joint optimization algorithm in this multi-tier computing network, which can be summarized as **Algorithm 1**.

Algorithm 1 The Overall Joint Optimization Algorithm in Multi-Tier Computing Networks

1: Initialize $\mathbf{A}^0$, $\mathbf{B}^0$, $\mathbf{P}^0$, $\mathbf{D}^0$, $\mathbf{T}^0$, $\mathbf{s}^0$, convergence threshold ϵ and iteration index $i = 0$.
2: **repeat**
3: Obtain subcarrier allocation $\mathbf{A}^*$, $\mathbf{B}^*$ and transmit power allocation $\mathbf{P}^*$ by solving the problem (P2).
4: Obtain task input bits $\mathbf{D}^*$, and time allocation $\mathbf{T}^*$, by solving the problem (P3.1).
5: Obtain RMS transmissive coefficient $\mathbf{s}^*$, by solving the problem (P4.3).
6: Update $i = i + 1$.
7: **until** The fractional decrease of the objective value is below a threshold ϵ.
8: **return** The subcarrier allocation, transmit power allocation, task input bits, time allocation and RMS transmissive coefficient design scheme.

E. Computational Complexity and Convergence Analysis

1) Computational Complexity Analysis: In each iteration, the problem (P2) is solved with the computational complexity of $\mathcal{O}\left((KN)^{3.5}\right)$, and the problem (P3.1) is solved with computational complexity of $\mathcal{O}\left(K^{3.5}\right)$ [42]. The problem (P4.3) solves a SDP problem by interior point method, so the computational complexity can be represented by $\mathcal{O}\left(M^{3.5}\right)$. It can be assumed that the number of iterations required for the algorithm to reach convergence is i, the computational complexity of the proposed algorithm can be expressed as $\mathcal{O}\left(i\left((KN)^{3.5} + K^{3.5} + M^{3.5}\right)\right)$.

2) Convergence Analysis: The convergence of the proposed joint subcarrier allocation, transmit power allocation, task input bits, time allocation and RMS transmissive coefficient optimization in multi-tier computing networks is elaborated as follows.

We define $\mathbf{A}^{(i)}$, $\mathbf{B}^{(i)}$, $\mathbf{P}^{(i)}$, $\mathbf{D}^{(i)}$, $\mathbf{T}^{(i)}$ and $\mathbf{s}^{(i)}$ as the i-th iteration solution of the problem (P2), (P3.1) and (P4.3). Herein, the objective function in the i-th iteration is denoted by $\mathcal{E}\left(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{P}^{(i)}, \mathbf{D}^{(i)}, \mathbf{T}^{(i)}, \mathbf{s}^{(i)}\right)$. In the step 3 of **Algorithm 1**, since subcarrier allocation and the transmit power allocation can be obtained for given $\mathbf{D}^{(i)}$, $\mathbf{T}^{(i)}$ and $\mathbf{s}^{(i)}$. Hence, we have

$$\begin{aligned}&\mathcal{E}\left(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{P}^{(i)}, \mathbf{D}^{(i)}, \mathbf{T}^{(i)}, \mathbf{s}^{(i)}\right)\\ &\qquad \geq \mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i)}, \mathbf{T}^{(i)}, \mathbf{s}^{(i)}\right).\end{aligned} \tag{54}$$

Similarly, in the step 4 of **Algorithm 1**, we can obtain the task input bits and time allocation when $\mathbf{A}^{(i+1)}$, $\mathbf{B}^{(i+1)}$, $\mathbf{P}^{(i+1)}$ and $\mathbf{s}^{(i)}$ are given. Herein, we also have

$$\begin{aligned}&\mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i)}, \mathbf{T}^{(i)}, \mathbf{s}^{(i)}\right)\\ &\quad \geq \mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i+1)}, \mathbf{T}^{(i+1)}, \mathbf{s}^{(i)}\right).\end{aligned} \tag{55}$$

In the step 5 of **Algorithm 1**, RMS transmissive coefficient can be obtained when $\mathbf{A}^{(i+1)}$, $\mathbf{B}^{(i+1)}$, $\mathbf{P}^{(i+1)}$, $\mathbf{D}^{(i+1)}$, $\mathbf{T}^{(i+1)}$ are given. Therefore, we have

$$\mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i+1)}, \mathbf{T}^{(i+1)}, \mathbf{s}^{(i)}\right)$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

$$\geq \mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i+1)}, \mathbf{T}^{(i+1)}, \mathbf{s}^{(i+1)}\right). \quad (56)$$

Based on the above, we can obtain

$$\mathcal{E}\left(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{P}^{(i)}, \mathbf{D}^{(i)}, \mathbf{T}^{(i)}, \mathbf{s}^{(i)}\right)$$
$$\geq \mathcal{E}\left(\mathbf{A}^{(i+1)}, \mathbf{B}^{(i+1)}, \mathbf{P}^{(i+1)}, \mathbf{D}^{(i+1)}, \mathbf{T}^{(i+1)}, \mathbf{s}^{(i+1)}\right). \quad (57)$$

which shows that the value of the objective function is non-increasing after each iteration of **Algorithm 1**. Since the objective function must be lower bounded by a finite value, the convergence of **Algorithm 1** can be guaranteed.

IV. Numerical Results

In this section, numerical results are provided to evaluate the effectiveness of the proposed joint optimization algorithm in multi-tier computing networks. We consider a three-dimensional (3D) coordinate system in this paper, where the RMS and the DF are located at (0m, 0m, 10m) and (25m, 25m, 10m), respectively, and $K = 5$ users are randomly and uniformly distributed in a circle whose origin is (0m, 0m, 0m) and a radius of 50m. Moreover, DF is equipped with single antenna, and RMS is equipped with $M = 25$ elements. The number of subcarriers is set to 20, i.e., $N = 20$. Assuming that the parameters of all users are the same, we set $W = 1\text{MHz}$, $\sigma^2 = -70\text{dBm}$, $c_t = c_r = c_m = 10^3$ cycles/bit, $\alpha_t = 10^{-27}$, $\alpha_r = 0.3 \times 10^{-27}$, $P^t_{\max} = P^r_{\max} = 40\text{dBm}$, $f_c = 3\text{GHz}$, $f_{t,\max} = 2\text{GHz}$, $f_{r,\max} = 3\text{GHz}$, and $f_{m,\max} = 5\text{GHz}$ in our numerical simulations [9]. Meanwhile, the path loss exponents is set as $\nu = \alpha = 3$. The path loss with a reference distance of 1m is set to $C_0 = -30\text{dB}$, and we set Rician factor to $\kappa_1 = \kappa_2 = 3\text{dB}$. In addition, the convergence threshold of the proposed algorithm is set to 10^{-3}.

We first evaluate the convergence of the proposed joint optimization algorithm. Fig. 3 illustrates the variation of total energy consumption with the number of iterations under different element numbers M. One can observe that the total energy consumption for all curves monotonically decreases as the number of iterations increases which means the proposed algorithm can quickly achieve convergence and has good convergence performance. In addition, as the number of RMS elements increases, lower energy consumes, which provides a scheme to reduce system energy consumption by increasing the number of transmissive RMS elements. Hence, the effectiveness and advantages of the transmissive RMS transceiver system are confirmed.

Then, We compare the performance of our proposed multi-tier computing model with other computing models as following: (1) **benchmark 1** (i.e., local computing): In this case, users execute all input task bits locally within the time duration T. (2) **benchmark 2** (i.e., partial offloading of computation collaboration): In this case, the task input bits are divided into two parts to be executed at the local and DF relays respectively. (3) **benchmark 3** (i.e., partial offloading of communication collaboration): In this case, tasks bits are assigned to users and RMS multi-antenna systems to execute, where DF relay assists in offloading. (4) **benchmark 4** (i.e., RMS-random-phase): In this case, based on the proposed algorithm, the phase of the RMS is not optimized, but a random phase is applied.

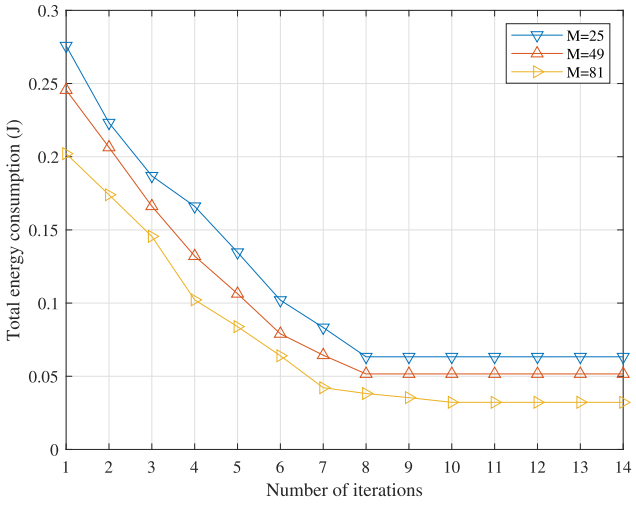


Fig. 3. Convergence behaviour of the proposed joint optimization algorithm.

We discuss the relationship between the computing capability of different benchmarks and the time duration T. Fig. 4 shows the number of input task bits versus the time duration T under different benchmarks. It can be observed that the number of task input bits under different benchmarks increases as the time duration T increases. That is because the longer time, the more tasks bits can be executed by the CPU. Specifically, the benchmark 2 and benchmark 3 can achieve higher computational capability than the benchmark 1. Furthermore, the benchmark 3 is superior to the benchmark 2 due to the high computation capability of the MEC server equipped with RMS transceiver. In plain sight, the proposed algorithm outperforms other benchmarks and achieves the highest computational capability by leveraging the strengths of both computation collaboration and communication collaboration. In addition, the proposed algorithm performs much better than the RMS-random phase benchmark due to the advantage of the optimized RMS transmissive coefficients.

Fig. 5 depicts the variation of the total energy consumption with the length of time duration T for the proposed algorithm and other benchmark algorithms. The total energy consumption for all the benchmraks decreases as time duration T increases. Specific observations are as follows. It can be seen that when the length of time duration T increases, the performance of the proposed algorithm improves and achieves the lowest total energy consumption. Since the transmit power is constant, the offloading time does not vary with T which results in the offloading energy consumption E_k^{off} remaining constant. However, as time increases, more task input bits can be offloaded to the MEC server, and the local computation energy consumption $E_k^{t,comp}$ given in Eq. (29) and computation energy consumption $E_k^{r,comp}$ given in Eq. (32) is subsequently reduced, resulting in a decreasing trend in total energy consumption. This demonstrates the benefits of MEC server for decreasing energy consumption. In addition, the benchmrak 4 performs worst due to the randomness of

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

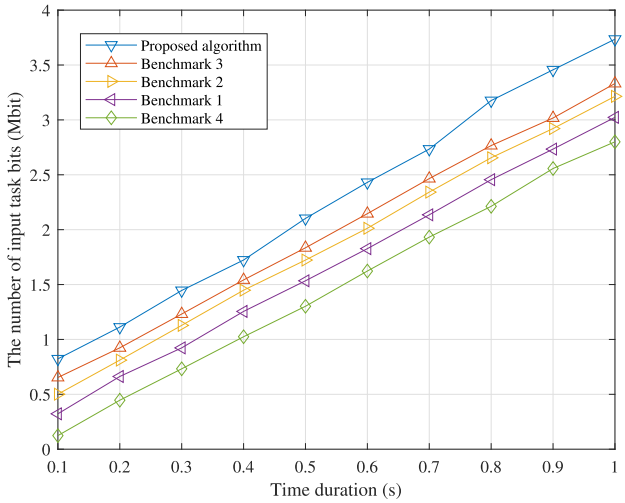


Fig. 4. Number of task input bits versus the time duration T for the proposed algorithm and different benchmark algorithms.

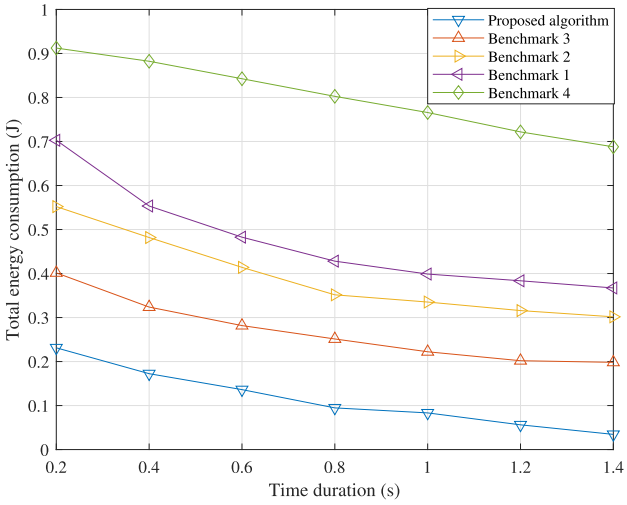


Fig. 5. Total energy consumption versus the time duration T for the proposed algorithm and different benchmark algorithms.

its phase. For the benchmark 1, as the time duration T increases, according to the expression given in Eq. (29), the local computation energy computation decreases. Meanwhile, both benchmark 2 and benchmark 3 outperform the local computation due to the exploit of computation resources at the DF relay and the transmissive RMS multi-antenna system.

Then, the variation of the total energy consumption with the task input bits D_k is shown in Fig. 6. The results show that the total energy consumption of the proposed algorithm and all benchmarks increases as the number of task input bits increases. This is because when the number of task input bits increases, the number of local and CN CPU operations increases, resulting in an increase in energy consumption. Furthermore, the proposed algorithm achieves minimum total energy consumption, especially for larger input bits, once again confirming the advantages of the proposed protocol in the multi-tier computing networks in terms of reducing total energy consumption.

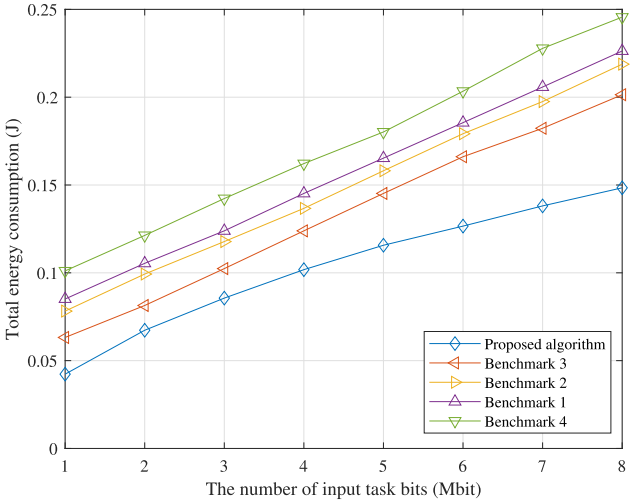


Fig. 6. Total energy consumption versus the task input bits for the proposed algorithm and different benchmark algorithms.

Next, Fig. 7 illustrates the variation of total energy consumption with the number of RMS transmissive elements. Obviously, it is not equipped with RMS transceiver in the benchmark 1 and benchmark 2, so the total energy consumption remains unchanged as the number of the RMS elements increases. We can observe that as the number of RMS elements increases, the total energy consumption decreases for the proposed algorithm and the benchmark 3 and benchmark 4, because the additional elements provide more channel diversity gain, which reduces the total energy consumption by reducing the transmit power. Specifically, when the number of RMS transmissive elements is the same, the performance of the proposed algorithm is better than the RMS-random-phase benchmark, for the reason that the optimized phase is controllable for the system and the effect of the random phase on the system is uncontrollable. In addition, the benchmark 3 performs less well than the proposed algorithm since lack of computing resources for DF relay. Meanwhile, the proposed algorithm achieves the lowest total energy consumption among all the benchmarks. Therefore, increasing the number of RMS transmissive elements is of great value for practical use.

Fig. 8 elaborates the variation of total energy consumption with the number of users. It can be seen from Fig. 9 that when the number of users increases, the total energy consumption increases. The reason is that as the number of users increases, the total input task bits increases, resulting in increased energy consumption for offloading and computation. When the number of users is the same, the proposed algorithm outperforms other benchmarks. The specific reasons are similar to those mentioned above, and are omitted here.

Afterwards, we evaluate the performance of our proposed joint optimization algorithm compared with other benchmark algorithms as follows. (4) **benchmark 4** (i.e., RMS-random-phase): In this case, based on the proposed algorithm, the phase of the RMS is not optimized, but a random phase is applied. (5) **benchmark 5** (i.e., RMS-SDR-phase): In this case, based on the proposed algorithm, the phase of the RMS is

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

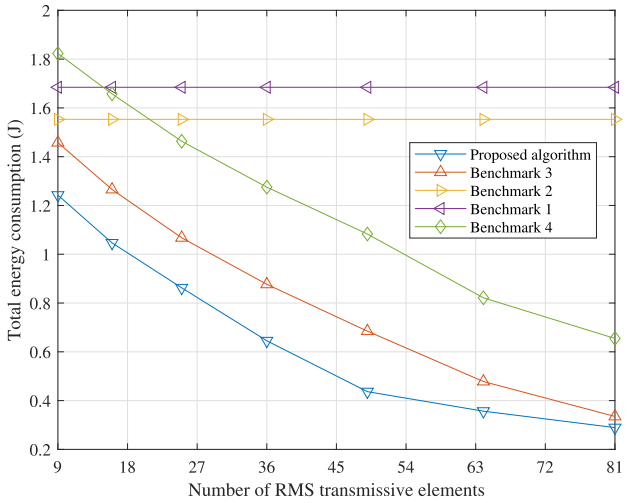


Fig. 7. Total energy consumption versus the number of RMS transmissive elements for the proposed algorithm and different benchmark algorithms.

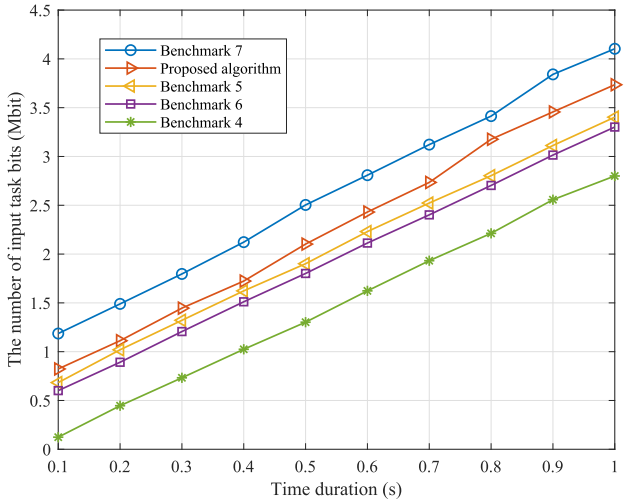


Fig. 9. Number of task input bits versus the time duration T for the proposed algorithm and different benchmark algorithms.

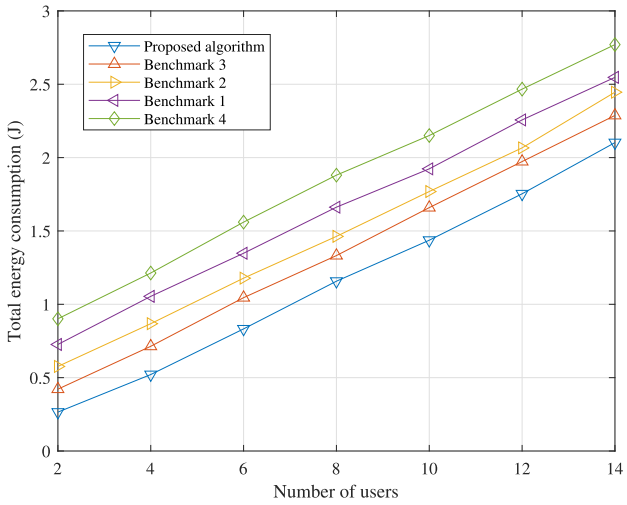


Fig. 8. Total energy consumption versus the number of users for the proposed algorithm and different benchmark algorithms.

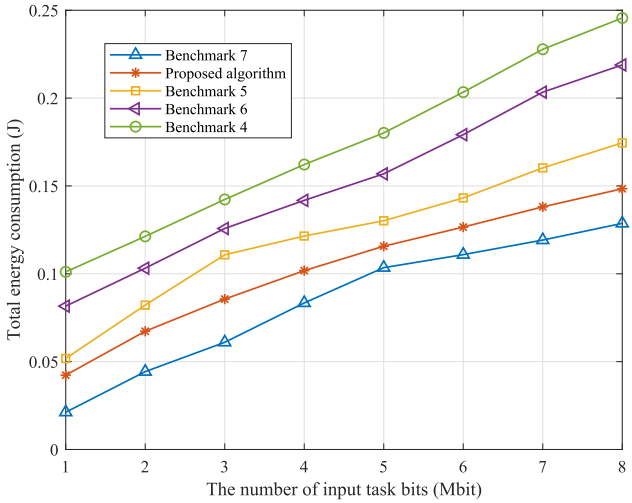


Fig. 10. Total energy consumption versus the task input bits for the proposed algorithm and different benchmark algorithms.

is optimized by implementing semidefinite relaxation (SDR) technique. (6) **benchmark 6** (i.e., three-stage algorithm): In this case, the three subproblems are optimized based on the proposed algorithm, but no alternating optimization is performed. (7) **benchmark 7** (i.e., upper bound): In this case, after solving sub-problem 1 to obtain the subcarrier assignment variable, it is not restored to a discrete binary variable.

Fig. 9 shows the variation of task input bits with time duration T under different optimization algorithms. It can be seen that, as the time duration T increases, since the CPU can execute more offloading tasks, the number of task input bits increases. In addition, within the same time duration T, the performance of benchmark 7 is better than our proposed algorithm and other benchmark algorithms. This is because after solving sub-problem 1, the benchmark algorithm does not approximately restore the sub-carrier allocation variables, which ensures that the algorithm performance. In addition, the performance of the proposed algorithm is better than benchmark 5 because the adopted DC algorithm has better performance guarantee than the SDR algorithm. Compared with the proposed algorithm, benchmark 6 does not perform alternating optimization after solving the three sub-problems, so it is easy to fall into local optimum and cannot achieve global optimum, so its performance is poor.

Finally, Fig. 10 illustrates the variation of total energy consumption with the number of task input bits under different optimization algorithms. It can be observed that when the number of task input bits increases, the total energy consumption also increases. When the number of task input bits is the same, the performance of benchmark 7 is better than that of the proposed algorithm and other benchmark algorithms, and the performance of the proposed algorithm is also better than that of other benchmark algorithms. The specific reasons are similar to the above, and are omitted here.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

This paper investigates the total energy consumption minimization problem of transmissive RMS transceiver enabled multi-tier computing networks. Specifically, under the constraints of the computing and energy resources, subcarrier allocation, input task bits allocation, time slot allocation, user and DF relay transmit power allocation, and the RMS transmissive coefficients are jointly optimized. First, we transform the problem into a tractable problem. Then, in order to solve the transformed problem, we apply the BCD algorithm framework to divide the original problem into three sub-problems for solving. Given the other variables, we solve the variables to be optimized through SCA, DC programming, etc., and then alternately optimize the three sub-problems until convergence is achieved. Then, the computational complexity and convergence analysis of the proposed algorithm are given. Finally, the numerical simulation results verify the convergence and effectiveness of the proposed algorithm, which illustrates that the proposed algorithm is capable of decreasing the total energy consumption to a large extent, and the advantages of utilizing transmissive RMS in this architecture for energy reduction are obvious.

References

[1] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[2] M. R. Palattella et al., "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[3] 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, Dec. 2017.

[4] X. Lyu et al., "Optimal schedule of mobile edge computing for Internet of Things using partial information," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2606–2615, Nov. 2017.

[5] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Energy-efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2603–2616, Jun. 2018.

[6] K. Wang, Y. Zhou, Q. Wu, W. Chen, and Y. Yang, "Task offloading in hybrid intelligent reflecting surface and massive MIMO relay networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3648–3663, Jun. 2022.

[7] H. A. Alameddine, S. Sharafeddine, S. Sebbah, S. Ayoubi, and C. Assi, "Dynamic task offloading and scheduling for low-latency IoT services in multi-access edge computing," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 668–682, Mar. 2019.

[8] K. Wang, W. Chen, J. Li, Y. Yang, and L. Hanzo, "Joint task offloading and caching for massive MIMO-aided multi-tier computing networks," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1820–1833, Mar. 2022.

[9] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4188–4200, Jun. 2019.

[10] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[11] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-efficient computation offloading for secure UAV-edge-computing systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6074–6087, Jun. 2019.

[12] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1757–1771, May 2016.

[13] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Aug. 2016.

[14] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5506–5519, Aug. 2018.

[15] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint computation offloading and user association in multi-task mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12313–12325, Dec. 2018.

[16] T. Ouyang, Z. Zhou, and X. Chen, "Follow me at the edge: Mobility-aware dynamic service placement for mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2333–2345, Oct. 2018.

[17] 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.

[18] 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.

[19] Z. Li, W. Chen, Q. Wu, K. Wang, and J. Li, "Joint beamforming design and power splitting optimization in IRS-assisted SWIPT NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 2019–2033, Mar. 2022.

[20] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4513–4526, Jul. 2021.

[21] H. Cao, Z. Li, and W. Chen, "Resource allocation for IRS-assisted wireless powered communication networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2450–2454, Nov. 2021.

[22] Z. Li, W. Chen, Q. Wu, H. Cao, K. Wang, and J. Li, "Robust beamforming design and time allocation for IRS-assisted wireless powered communication networks," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2838–2852, Apr. 2022.

[23] W. Yan, X. Yuan, Z.-Q. He, and X. Kuai, "Passive beamforming and information transfer design for reconfigurable intelligent surfaces aided multiuser MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1793–1808, Aug. 2020.

[24] 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.

[25] H. Ur Rehman, F. Bellili, A. Mezghani, and E. Hossain, "Joint active and passive beamforming design for IRS-assisted multi-user MIMO systems: A VAMP-based approach," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6734–6749, Oct. 2021.

[26] B. Zheng and R. Zhang, "IRS meets relaying: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 2080–2084, Sep. 2021.

[27] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[28] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021.

[29] S. Zeng et al., "Reconfigurable intelligent surfaces in 6G: Reflective, transmissive, or both?" *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 2063–2067, Jun. 2021.

[30] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, "Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13905–13909, Nov. 2020.

[31] S. Zhang et al., "Intelligent omni-surfaces: Ubiquitous wireless transmission by reflective-refractive metasurfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 219–233, Jan. 2022.

[32] Y. Liu et al., "Star: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 102–109, Dec. 2021.

[33] H. Niu, Z. Chu, F. Zhou, P. Xiao, and N. Al-Dhahir, "Weighted sum rate optimization for STAR-RIS-assisted MIMO system," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2122–2127, Feb. 2022.

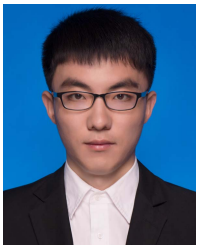
[34] W. Tang et al., "MIMO transmission through reconfigurable intelligent surface: System design, analysis, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2683–2699, Nov. 2020.

[35] X. Bai et al., "High-efficiency transmissive programmable metasurface for multimode OAM generation," *Adv. Opt. Mater.*, vol. 8, no. 17, Sep. 2020, Art. no. 2000570.

[36] Z. Li, W. Chen, and H. Cao, "Beamforming design and power allocation for transmissive RMS-based transmitter architectures," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 53–57, Jan. 2022.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.

[37] Z. Li, W. Chen, Q. Wu, J. Lu, K. Wang, and J. Li, "Uplink transceiver design and optimization for transmissive RMS multi-antenna systems," 2021, *arXiv:2112.08880*.
[38] H. Liu, J. Zhang, Q. Wu, H. Xiao, and B. Ai, "ADMM based channel estimation for RISs aided millimeter wave communications," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 2894–2898, Sep. 2021.
[39] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.
[40] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.
[41] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 2.1," Mar. 2014. [Online]. Available: http://cvxr.com/cvx/
[42] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.



Zhendong Li received the B.S. degree in communications engineering from Zhengzhou University in 2017 and the master's degree in telecommunication and information systems from the Beijing University of Posts and Telecommunications in 2020. He is currently pursuing the Ph.D. degree with the Broadband Access Network Laboratory, Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China. His research interests include reconfigurable meta-surface (RMS), unmanned aerial vehicle (UAV) communications, space-air-ground (SAG) networks, the Internet of Things (IoT), and wireless resource management in future wireless networks.

Wen Chen (Senior Member, IEEE) is currently a Tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is also the Director of the Broadband Access Network Laboratory. He is a fellow of Chinese Institute of Electronics and the 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. His research interests include multiple access, wireless AI, and meta-surface communications. He has published more than 120 papers in IEEE journals and more than 120 papers in IEEE conferences, with citations more than 8000 in Google scholar.

Ziwei Liu received the B.S. degree in electronic and information engineering from Northwestern Polytechnical University in 2021. He is currently pursuing the Ph.D. degree with the Broadband Access Network Laboratory, Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China. His research interests include reconfigurable intelligent surface (RIS), random access, and wireless resource management in future wireless networks.

Hongying Tang received the Ph.D. degree from Shanghai Jiao Tong University (SJTU), China, in 2015. She is currently a Senior Engineer with the Science and Technology on Microsystem Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences. Her research interests include unmanned aerial vehicle (UAV) communications and MAC protocols in wireless sensor networks.

Jianmin Lu (Member, IEEE) joined the Huawei Technologies in 1999. During the last two decades, he conducted various researches on wireless communications, especially on physic layer and MAC layer and developed 3G, 4G, and 5G products. He is currently an Executive Director of Huawei Wireless Technology Laboratory. He received more than 50 patents during the research. He was deeply involved in 3GPP2 (EVDO/UMB), WiMAX/802.16m, and 3GPP (LTE/NR) standardization and contributed several key technologies, such as flexible radio frame structure, radio resource management, and MIMO. His current research interests include the area of signal processing, protocol, and networking for the next generation wireless communication.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 27,2023 at 05:21:12 UTC from IEEE Xplore. Restrictions apply.