Sum Rate Maximization Under AoI Constraints for RIS-Assisted mmWave Communications

Ziqi Guo, Yong Niu, *Senior Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, Changming Zhang, Ning Wang, *Member, IEEE*, Zhangdui Zhong, *Fellow, IEEE*, and Bo Ai, *Fellow, IEEE*

***Abstract*—The concept of age of information (AoI) has been proposed to quantify information freshness, which is crucial for time-sensitive applications. However, in millimeter wave (mmWave) communication systems, the link blockage caused by obstacles and the severe path loss greatly impair the freshness of information received by the user equipments (UEs). In this article, we focus on reconfigurable intelligent surface (RIS)-assisted mmWave communications, where beamforming is performed at transceivers to provide directional beam gain and a RIS is deployed to combat link blockage. We aim to maximize the system sum rate while satisfying the information freshness requirements of UEs by jointly optimizing the beamforming at transceivers, the discrete RIS reflection coefficients, and the UE scheduling strategy. To facilitate a practical solution, we decompose the problem into two subproblems. For the first per-UE data rate maximization problem, we further decompose it into a beamforming optimization subproblem and a RIS reflection coefficient optimization subproblem. Considering the difficulty of channel estimation, we utilize the hierarchical search method for the former and the local search method for the latter, and then adopt the block coordinate descent (BCD) method to alternately solve them. For the second scheduling strategy design problem, a low-complexity heuristic scheduling algorithm is designed. Simulation results show that the proposed algorithm can effectively improve the system sum rate while satisfying the information freshness requirements of all UEs.**

Manuscript received 1 March 2023; revised 29 August 2023; accepted 6 November 2023. Date of publication 10 November 2023; date of current version 22 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2900301, in part by the National Key Research and Development Program of China under Grant 2020YFB1806903, in part by the National Natural Science Foundation of China under Grants 62221001, 62231009, and U21A20445, in part by the Fundamental Research Funds for the Central Universities, China, under Grants 2022JBQY004 and 2022JBXT001, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023JBMC030. The review of this article was coordinated by Dr. Benedetta Picano. *(Corresponding authors: Yong Niu; Bo Ai.)*

Ziqi Guo is with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China, and also with the Collaborative Innovation Center of Railway Traffic Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: 21120053@bjtu.edu.cn).

Yong Niu is with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China (e-mail: niuy11@163.com).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: smao@ieee.org).

Changming Zhang is with the Research Institute of Intelligent Networks, Zhejiang Lab, Hangzhou 311121, China (e-mail: zhangcm@zhejianglab.com).

Ning Wang is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: ienwang@zzu.edu.cn).

Zhangdui Zhong and Bo Ai are with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications, Beijing Jiaotong University, Beijing 100044, China (e-mail: zhdzhong@bjtu.edu.cn; aibo@ieee.org).

Digital Object Identifier 10.1109/TVT.2023.3331707

***Index Terms*—Reconfigurable intelligent surface (RIS), age of information (AoI), beamforming, discrete phase shifts, scheduling.**

I. Introduction

In recent years, a variety of novel applications have emerged, leading to a dramatic increase in mobile data traffic. According to the International Telecommunication Union (ITU), mobile data traffic is predicted to grow from 62 EB per month in 2020 to 5,016 EB per month in 2030 [1]. This puts a compelling need for higher capacity of communication systems and further intensifies the conflict between the demands for communication capacity and the scarce spectrum resources. Therefore, millimeter wave (mmWave) communication is considered as a promising technology for future cellular networks due to its large available bandwidth [2], [3], [4].

On the other hand, many new time-sensitive applications, e.g., autonomous driving, depend on timely and reliable information exchange. Once information is generated, it should be sent to the receiver for timely processing, and outdated information could seriously degrade the user's experience. In order to capture the information freshness, age of information (AoI) has been proposed, which is defined as the elapsed time since the generation of the most recently received status-update [5], [6]. Receivers wish to receive data with a lower AoI, so that the received data will be fresher. The data received with a high AoI could be meaningless or harmful.

In practical communication systems, AoI is generally affected by the scheduling strategy and the quality of the received signal. However, compared with microwave communication below 6 GHz, a key challenge of mmWave communication is that the signal in the mmWave band will experience more severe path loss due to the short wavelength [7], which degrades the quality of the received signal. It is necessary to establish a directional transmission link between transceivers with the help of large-scale antenna arrays and beamforming, which can provide high antenna gains for mmWave signals to compensate for path loss. However, the directional transmissions and weak diffraction ability make mmWave signals vulnerable to blockage, especially in indoor and dense urban environments [8]. The high AoI due to the blockage nature is often unacceptable in most time-sensitive applications.

Fortunately, reconfigurable intelligent surface (RIS) can flexibly configure the propagation environment through software programming, which can be used to combat mmWave link blockage [9]. Specifically, RIS is a device composed of a large number

of passive reconfigurable reflection elements. Each element can independently control the amplitude and phase changes to the incident signal in a software-defined manner [10]. By a proper design, the passive reflections of all the reflection elements of RIS can be coherently superposed at the desired receiver to increase the received signal power, thus creating a more reliable reflection link and avoiding blockage of the signal in the direct link. Therefore, deploying RIS in mmWave systems exhibits the potential to achieve superior information freshness performance. However, as a passive reflective device, RIS is not capable of transmitting, processing, and receiving signals. The task of channel estimation grows increasingly challenging as the count of reflection elements escalates. Besides, compared with other wireless systems, the large-scale antenna arrays in mmWave systems greatly increase the difficulty of channel estimation [11]. Therefore, it is necessary to discuss how to guarantee the information freshness performance of the system without knowing CSI.

Therefore, in this article, we study a downlink RIS-assisted mmWave MIMO system, where the base station (BS) transmits time-sensitive data to user equipments (UEs). Considering the difficulty of channel estimation in the system, we assume that the full CSI is unknown. Different from most of the existing AoI research on the overall AoI minimization, we wish to satisfy the information freshness requirement of each UE in the system, which can provide a better communication experience for UEs. Besides, we pursue the maximization of the system sum rate while satisfying the information freshness requirements, which can further stimulate the potential of mmWave communication systems. Thus, our work aims to maximize the system sum rate over a fixed time interval, i.e., a superframe, while satisfying all the UEs' information freshness requirements in the system. Optimization variables include the beamforming vectors at the BS and UEs, the discrete RIS reflection coefficients, and the scheduling matrix, all of which are coupled in the expressions of the sum rate in the objective function and AoI constraints. The problem is an integer non-convex optimization problem. To reduce the complexity of the solution, we decompose the optimization problem into several subproblems and solve them separately.

The contributions of this article are summarized as follows:

- We study a downlink RIS-assisted mmWave MIMO system, in which a RIS is deployed to provide a reliable reflection path against the blockage of direct links, and time-sensitive data is transmitted from the BS to UEs. Each UE in the system has certain requirement for the freshness of information.
- We formulate the system sum rate maximization problem by optimizing the beamforming vectors at the BS and UEs, the RIS reflection coefficients, and the scheduling matrix, subject to the AoI constraints of UEs. Since all the optimization variables are coupled, the complexity of finding the overall optimal solution by exhaustive search will be prohibitively high. To address this issue, we decompose the original problem into a per-UE rate maximization problem and a scheduling strategy design problem.
- For the per-UE rate maximization problem, considering that the full CSI is unknown, the hierarchical search method and local search method are used for the optimization of the beamforming vectors and the RIS reflection coefficients, respectively. Due to the coupled beamforming vectors and RIS reflection coefficients, we use the block coordinate descent (BCD) algorithm to iteratively update the two sets of optimization variables. For the scheduling strategy design problem, we propose a low-complexity heuristic strategy, which maximizes the system sum rate over a superframe while satisfying the AoI constraints.
- We evaluate the performance of the proposed algorithm with simulations. Compared with three benchmark schemes, the simulation results demonstrate that the proposed algorithm ensures the information freshness requirements of all UEs, and the system sum rate is effectively improved.

The rest of the article is organized as follows. Section II reviews related work. The system overview and problem formulation are presented in Sections III and IV, respectively. We present the sum rate maximization algorithm in Section V and discuss our simulation results in Section VI. Section VII concludes this article.

II. Related Work

In RIS-assisted communication systems, a key problem of interest is to jointly devise the RIS reflection coefficients and the active beamforming vectors at the BS to improve system performance. Numerous studies have been conducted to solve this problem under different system setups and assumptions, some of which are for mmWave MIMO communication systems. Perovic et al. [12] compared two optimization schemes in the indoor RIS-assisted mmWave environment without the line-of-sight (LOS) path. They showed the joint optimization of the RIS reflection elements and the transmit phase precoder can effectively enhance channel capacity. Wang et al. [13] considered a RIS-assisted downlink mmWave system with a hybrid beamforming structure. A manifold optimization (MO)-based algorithm was developed to jointly optimize the RIS's reflection coefficients and the hybrid beamforming at the BS for maximization of spectral efficiency. Feng et al. [14] designed a successive interference cancelation (SIC)-based method for the bandwidth-efficiency maximization problem. A greedy method is proposed for the hybrid beamforming design and a complex circle manifold (CCM)-based method is used for updating of the RIS elements. Li et al. [15] formulated a power minimization problem with signal-to-interference-plus-noise ratio (SINR) constraints in multi-user scenarios. They proposed a two-layer penalty-based algorithm to decouple variables in SINR constraints and three different methods to optimize the BS analog beamforming and the RIS response matrix in the penalty-based algorithm.

The above works rely on full CSI through channel estimation. Considering the difficulty of channel estimation in RIS-assisted mmWave MIMO systems, some studies focus on the problem of

beam training with the goal of obtaining the angle of departure (AoD) and angle of arrival (AoA) associated with the dominant path. Wang et al. [16] developed an efficient downlink beam training method for RIS-assisted mmWave or THz systems. They designed multi-directional beam training sequences to scan the angular space and proposed an efficient set-intersection-based scheme to identify the best beam alignment. Wei et al. [17] proposed an effective near-field beam training scheme by designing a near-field codebook that matches the near-field channel model for the extremely large-scale RIS-assisted system. Wang et al. [18] considered a multi-RIS-assisted mmWave MIMO system and carried out beam training designs with random beamforming and maximum likelihood (ML) estimation to estimate the parameters of the LOS component.

Recently, the importance of information freshness has been recognized and AoI has been considered in the design of wireless communication systems. For example, He et al. [19] designed two scheduling strategies for the system AoI minimization problem in wireless networks. One is based on the ILP formulation, and the other is the suboptimal but more scalable steepest age descent algorithm. Kadota et al. [20] formulated the problem of minimizing the expected weighted sum of AoI under time-throughput constraints. They designed four low-complexity scheduling strategies for solving this problem and found the Max-Weight and the Drift-Plus-Penalty have better performance in terms of AoI and throughput. Liu et al. [21] proved that any optimal solution of the maximum delay minimization problem is an approximate solution of the AoI minimization problem with bounded optimality loss. Inspired by this, a framework was developed to solve the AoI minimization problem in multi-path communications. Bhat et al. [22] considered the long-term average throughput maximization problem in fading channels, where the system average AoI and power are regarded as constraints. They proposed a simple age-independent stationary randomized power allocation policy to solve the problem. In addition, the optimization of AoI has been extended to different application scenarios, such as multi-access edge computing-assisted IoT networks [23], unmanned aerial vehicles (UAV) communications [24], simultaneous wireless information and power transfer (SWIPT) enabled communications [25], and the joint radar-communication (JRC) [26], etc.

There were also some related works on AoI optimization in RIS-assisted wireless communications [27], [28], [29], [30], [31], [32], [33]. Sorkhoh et al. [27] studied the RIS-assisted cooperative autonomous driving (CAD) systems. They scheduled the resource blocks and RISs to minimize the average AoI of all streams. Muhammad et al. [28] examined the joint optimization of the RIS phase shifts and the traffic streams scheduling based on semi-definite relaxation (SDR), and solved the problem of minimizing the sum AoI in RIS-assisted wireless networks in single-antenna scenarios. Samir et al. [29] formulated an optimization problem with the objective of minimizing the expected sum AoI in an IoT network with the relay of a UAV equipped with RIS. To solve this problem, they developed a deep reinforcement learning (DRL) framework to jointly optimize the UAV height, RIS phase shift, and scheduling strategy. Fan et al. [30] deployed a RIS between IoT devices and UAVs to overcome the obstacles of urban buildings, and designs a DRL scheme to optimize UAV trajectory, discrete RIS phase shift, and scheduling strategy to minimize the total AoI of all devices. Feng et al. [31] adopted the DRL algorithm to jointly optimize the phase-shift matrix of RIS and service time of packets to solve the problem of minimizing the average peak information age in RIS-assisted non-orthogonal multiple access (NOMA) networks. Lyu et al. [32] investigated the sum AoI minimization in a RIS-assisted SWIPT network, where the energy harvesting demands of users were considered. They proposed a successive convex approximation (SCA) based alternating optimization (AO) algorithm to handle the scheduling problem with joint active and passive beamforming design. Shi et al. [33] considered the average AoI minimization through the joint design of the RIS phase shifts, transmit powers, and transmission rate in hybrid automatic repeat request (HARQ)-RIS aided IoT networks. However, all of these studies took the overall AoI minimization as their objective and neglected the information freshness requirements of UEs. Moreover, the AoI optimization in RIS-assisted mmWave communications has not been studied yet. Therefore, in this article, we focus on sum rate maximization while satisfying the information freshness requirements of UEs in RIS-assisted mmWave MIMO systems.

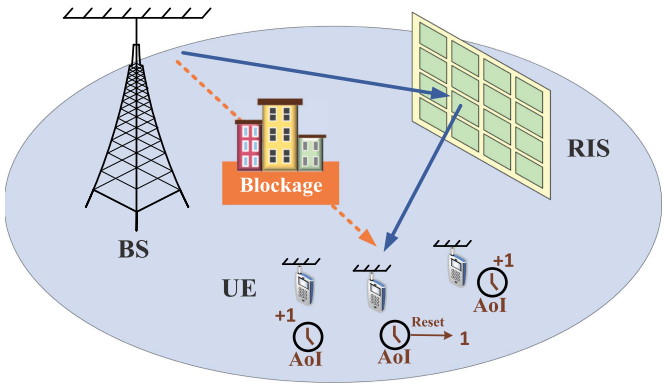

Fig. 1. Illustration of the system model.

III. System Overview

A. System Model

As shown in Fig. 1, we consider a single-cell mmWave MIMO communication system, where multiple UEs need to obtain fresh data from the BS. A typical example is that UEs require real-time traffic information from the BS for trip planning. The set of UEs is denoted as $\mathcal{K} = \{1, 2, \ldots, K\}$. The BS is equipped with N_t antennas and each UE is equipped with N_r antennas. Both the BS and the UEs use the uniform linear array (ULA) antennas. The direct links between the BS and the UEs are assumed to be blocked by some obstacles, e.g., high buildings. Thus, a RIS with M passive reflection elements is deployed to provide a reliable reflection link for the UEs.

Moreover, time is divided into a series of non-overlapping superframes. Each superframe consists of two phases: the scheduling phase and the transmission phase. In the scheduling phase, the scheduling and network optimization scheme is computed

by a central controller located at the BS, then the results are sent to the RIS and the UEs. In the transmission phase, the BS communicates with the UEs with the help of RIS following the scheme. If the system changes during the transmission phase causing a transmission failure, UEs will report the transmission failure to the BS. The system will advance to the next superframe and the BS will redesign the scheme.

We focus on the performance of the system in the transmission phase. The transmission phase can be divided equally into T time slots, denoted as $\mathcal{T}=\{1,2,\ldots,T\}$. The time-division multiple access (TDMA) protocol is adopted, which means that only one UE can be scheduled for each time slot, so there is no interference between different UEs [30], [34]. At each time slot, the BS side and the scheduled UE side perform analog beamforming to generate directional antenna gain. We assume that for each time slot, the BS transmits a signal with the same power P_T. The signal transmitted by the BS in time slot t can be expressed as $\mathbf{x}_t=\mathbf{w}_t\sqrt{P_T}s_t$, where s_t denotes the transmitted data at time slot t with $\mathbb{E}\{s_t\}=0$ and $\mathbb{E}\{s_t s_t^H\}=1$, and $\mathbf{w}_t\in\mathbb{C}^{N_t\times 1}$ denotes corresponding beamforming vector at the BS. The received signal at the scheduled UE in time slot t is expressed as

$$y_t=\mathbf{f}_t^H\left(\mathbf{H}_t^H\mathbf{x}_t+\mathbf{n}_t\right)=\mathbf{f}_t^H\left(\mathbf{H}_t^H\mathbf{w}_t\sqrt{P_T}s_t+\mathbf{n}_t\right), \quad (1)$$

where $\mathbf{f}_t\in\mathbb{C}^{N_r\times 1}$ denotes the beamforming vector at the UE, $\mathbf{n}_t\sim\mathcal{CN}(\mathbf{0},\sigma^2\mathbf{I}_{N_r})$ is the additive Gaussian white noise received by the UE, and $\mathbf{H}_t\in\mathbb{C}^{N_t\times N_r}$ denotes the channel matrix in time slot t.

Since the direct link is blocked, the transmitted signal arrives at the UE via the BS-RIS-UE channel. The RIS is a uniform planar array (UPA) consisting of M passive reflection elements, each of which can independently adjust the amplitude and phase of the incident signal. In view of the severe path loss, we ignore the signals reflected by the RIS twice and more and consider only the signal reflected for the first time [35]. Let $\mathbf{G}_t\in\mathbb{C}^{N_t\times M}$ and $\mathbf{H}_{r,t}\in\mathbb{C}^{M\times N_r}$ represent the reflection channel matrixes at time slot t from the BS to the RIS and from the RIS to the UE, respectively. Thus, the channel matrix $\mathbf{H}_t$ can be expressed as

$$\mathbf{H}_t=\mathbf{G}_t\mathbf{\Phi}_t\mathbf{H}_{r,t}. \quad (2)$$

Here, $\mathbf{\Phi}_t=\mathrm{diag}(\beta_{1,t}e^{j\varphi_{1,t}},\ \beta_{2,t}e^{j\varphi_{2,t}},\ \cdots\ \beta_{M,t}e^{j\varphi_{M,t}})\in\mathbb{C}^{M\times M}$ denotes the reflection-coefficient matrix of the RIS, where $\beta_{m,t}\in[0,1]$ and $\varphi_{m,t}\in[0,2\pi]$ represent the amplitude reflection coefficient and phase-shift reflection coefficient of RIS element m in time slot t, respectively. For simplicity, each reflection element of RIS is designed to maximize signal reflection (i.e., $\beta_{m,t}=1,\ \forall m,t$) [36], [37]. Further, for the sake of hardware implementation, the phase shift of RIS takes finite discrete values. We assume that each RIS element can realize 2^b different discrete phase shift values by b-bit quantization, the set of discrete phase shifts is represented as $\mathcal{F}=\{0,\frac{2\pi}{2^b},\ldots,(2^b-1)\frac{2\pi}{2^b-1}\}$ [38].

Accordingly, the SNR received by the UE scheduled in time slot t is given by

$$\gamma_t=\frac{\left|\mathbf{f}_t^H(\mathbf{G}_t\mathbf{\Phi}_t\mathbf{H}_{r,t})^H\mathbf{w}_t\sqrt{P_T}\right|^2}{\sigma^2}. \quad (3)$$

To ensure that the UE can correctly demodulate the signal, the SNR should be greater than a threshold value γ_{th}, i.e., $\gamma_t>\gamma_{th}$. Then, the achievable transmission rate in time slot t can be written as

$$R_t=\log_2(1+\gamma_t). \quad (4)$$

However, it is worth noting that obtaining the full CSI by channel estimation is difficult in this system [11]. Moreover, the RIS in the system is a passive device with no RF chain. Therefore, it cannot receive, transmit, and process signals other than just reflecting signals. It cannot directly estimate the BS-RIS channel and the RIS-UE channel. On the other hand, the large antenna array and the large number of passive reflection elements of RIS impose a substantial overhead on channel estimation. This is very detrimental to the design and optimization of the system. Therefore, we adopt the beam training method in the scheduling phase to obtain the AoD and AoA associated with the reflection path, instead of explicitly estimating the entire channel. Specially, the beam search space is represented by a codebook containing multiple codewords. We denote the codebook of the BS and each UE as Γ_t and Γ_r, respectively. Thus, we have $\mathbf{w}_t\in\Gamma_t$ and $\mathbf{f}_t\in\Gamma_r$ for $\forall t$. In the scheduling phase, the BS consecutively sends beam training signals to each UE through the reflection of the RIS. Both the BS and each UE can sweep the beamforming vectors in the pre-designed codebook, while the different phase shift of each RIS element is selected from $\mathcal{F}$ to change the reflection beam direction. Then, based on the UE's feedback, the combination of beamforming vectors and RIS reflection coefficients that maximizes the UE's achievable transmission rate will be selected.

B. Channel Model

Due to the small wavelength, mmWave signals exhibit weak diffraction capabilities and suffer from high path loss, which makes the mmWave channel have limited scattering. The number of scatterers is typically substantially fewer than the number of antennas at the transceiver. Moreover, the dense configurations of antenna arrays in mmWave transceivers introduce pronounced antenna correlation. Given this, the Saleh-Valenzuela (S-V) channel model has been extensively used the capture the mathematical attributes of the mmWave channels [39]. In Fig. 2, we present a schematic diagram of the BS-RIS channel based on the S-V channel model. The channel matrix between transceivers can be portrayed as a superposition of multipath components, where different multipath components have different angles of separation (AoDs) and angles of arrival (AoAs).

Assume that the channels do not change within a superframe. In each time slot t, the BS-RIS channel $\mathbf{G}_t$ and the RIS-UE channel $\mathbf{H}_{r,t}$ can be written as

$$\mathbf{G}_t=\sqrt{\frac{N_tM}{P}}\sum_{i=1}^{P}\tilde{\alpha}_i\mathbf{a}_r\left(M,\phi_{RIS,i}^r,\zeta_{RIS,i}^r\right)\mathbf{a}_t^H\left(N_t,\psi_{BS,i}^t\right), \quad (5)$$

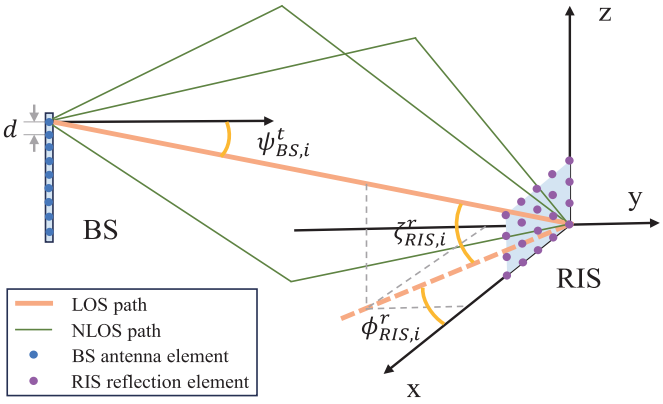

Fig. 2. Illustration of the BS-RIS channel model.

$$\mathbf{H}_{r,t} = \sqrt{\frac{MN_r}{L}} \sum_{i=1}^{L} \tilde{\beta}_i \mathbf{a}_r\left(N_r, \psi^r_{UE,i}\right) \mathbf{a}_t^H\left(M, \phi^t_{RIS,i}, \zeta^t_{RIS,i}\right), \tag{6}$$

where P is the total number of paths between the BS and the RIS, L is the total number of paths between the RIS and the UE scheduled in time slot t. $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ denote the complex gain of the i-th path. $\phi^r_{RIS,i}$ and $\zeta^r_{RIS,i}$ represent the azimuth and elevation angles of arrival associated with the RIS, respectively, while $\phi^t_{RIS,i}$ and $\zeta^t_{RIS,i}$ represent the azimuth and elevation angles of departure associated with the RIS, respectively. $\psi^t_{BS,i}$ denotes the angle of departure from the BS and $\psi^r_{UE,i}$ denotes the angle of arrival to the scheduled UE. $\mathbf{a}_r(\cdot)$ and $\mathbf{a}_t(\cdot)$ denote the normalized angle steering vector functions at transmitter and receiver, respectively. Specifically, for the BS and UEs with an N-element ULA, the corresponding angle steering vector is expressed as

$$\mathbf{a}(N,\psi) = \frac{1}{\sqrt{N}}\left[1, e^{j\frac{2\pi d}{\lambda}\sin(\psi)}, \ldots, e^{j\frac{2\pi d}{\lambda}(N-1)\sin(\psi)}\right], \tag{7}$$

and for the RIS with the UPA with $M = M_a \times M_b$ reflection elements, the corresponding normalized angle steering vector is expressed as

$$\mathbf{a}(M,\phi,\zeta) = \frac{1}{\sqrt{M}}\Big[1, \ldots, \mathrm{e}^{j\frac{2\pi d}{\lambda}((m_a-1)\sin(\zeta)\sin(\phi)+(m_b-1)\cos(\zeta))} \cdots, \mathrm{e}^{j\frac{2\pi d}{\lambda}((M_a-1)\sin(\zeta)\sin(\phi)+(M_b-1)\cos(\zeta))}\Big]. \tag{8}$$

C. AoI Definition

We use the AoI to measure the freshness of information received by UEs. The BS is assumed to follow a *per time slot sampling strategy*, i.e., it samples status-update information and sends a status-update packet at the beginning of each time slot [29]. Meanwhile, a single packet queue discipline is considered for the BS, which means the older status-update packet will be replaced by a newly arrived packet. We use $u_{k,t} \in \{0,1\}$ to indicate whether UE k is scheduled to receive data from the BS in time slot t. If UE k is scheduled, $u_{k,t} = 1$, and then the BS sends a status-update packet to UE k; otherwise, $u_{k,t} = 0$. The overall scheduling strategy in time slot t is expressed as $\mathbf{u}_t = [u_{1,t}, u_{2,t}, \ldots, u_{K,t}]^T$. Note that in addition to being scheduled, the successful transmission of the status-update information to UE k requires that the SNR exceeds the threshold for reliable demodulation. In case the status-update packet is successfully transmitted to UE k in time slot t, the AoI of UE k is reset to 1, otherwise, the AoI is increased by 1. Therefore, the evolution of the AoI of UE k is given by

$$\mathcal{A}_{k,t} = \begin{cases} 1, & \text{if } u_{k,t} = 1 \text{ and } \gamma_t > \gamma_{th}, \\ \mathcal{A}_{k,t-1} + 1, & \text{otherwise.} \end{cases} \tag{9}$$

For simplicity, we assume that the initial AoI $\mathcal{A}_{k,0} = 1, \forall k$. The average AoI of UE k during the transmission phase consisting of T time slots is given by

$$\mathcal{A}_k = \frac{1}{T}\sum_{t=1}^{T} \mathcal{A}_{k,t}. \tag{10}$$

Considering the requirement of each UE for fresh information, we denote the maximum tolerable AoI for UE k as $\mathcal{A}_{k,\max}$. The AoI of each UE should satisfy $\mathcal{A}_k \le \mathcal{A}_{k,\max}, \forall k$. In this article, we focus on the situation in which each UE receives the same types of service from the BS. Generally, we assume $\mathcal{A}_{k,\max} = \mathcal{A}_{\max}, \forall k$.

IV. Problem Formulation and Decomposition

In this section, we first formulate our optimization problem based on the above system model, and then decompose the complex problem in order to efficiently solve it.

A. Sum Rate Maximization Problem Formulation

In this article, we aim to maximize the sum rate of the system over T time slots by jointly optimizing the scheduling strategy, the reflection-coefficient matrix of the RIS, and the beamforming vector of the BS and UEs. To facilitate the subsequent presentation, let $\mathbf{\Phi} = [\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_T]^T$ denote the RIS reflection coefficient matrix over T time slots, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T]^T$ and $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T]^T$ denote the beamforming vector of the BS and UEs over T time slots, respectively, and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T]^T$ represent the scheduling strategy over T time slots. Then, the joint optimization problem (P1) can be formulated as

$$\max_{\mathbf{U},\mathbf{W},\mathbf{\Phi},\mathbf{F}} \sum_{t=1}^{T} R_t \tag{11}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} u_{k,t} = 1, \ \forall t \in \{1,2,\ldots,T\}, \tag{12}$$

$$\mathcal{A}_k \le \mathcal{A}_{k,\max}, \ \forall k \in \{1,2,\ldots,K\}, \tag{13}$$

$$u_{k,t} \in \{0,1\}, \forall k \in \{1,2,\ldots,K\}, t \in \{1,2,\ldots,T\}, \tag{14}$$

$$\varphi_{m,t} \in \mathcal{F}, \ \forall m \in \{1,2,\ldots,M\}, t \in \{1,2,\ldots,T\}, \tag{15}$$

$$\mathbf{w}_t \in \Gamma_t, \ \forall t \in \{1,2,\ldots,T\}, \tag{16}$$

$$\mathbf{f}_t \in \Gamma_r, \ \forall t \in \{1,2,\ldots,T\}, \tag{17}$$

where Constraint (12) indicates that only one UE is scheduled in each time slot, and Constraint (13) guarantees that the information freshness of each UE is ensured. Constraint (14) limits scheduling variables $u_{k,t}$ to 0-1 variables, and Constraints (15)–(17) restrict $\varphi_{m,t}$, $\mathbf{w}_t$, and $\mathbf{f}_t$ to be discrete values. The problem is an integer non-convex optimization problem. Moreover, the four variables, $\mathbf{U}$, $\mathbf{W}$, $\mathbf{\Phi}$, and $\mathbf{F}$, are coupled in both the objective function and $\mathcal{A}_k$. Although the global optimal solution can be found by exhaustive search, the multi-variable coupling makes the search space prohibitively large and consequently, the computational overhead considerable. Facing these challenges, our goal is to design a low-complexity algorithm to solve this problem.

B. Problem Decomposition

Considering the multi-variable coupling, we first decompose the problem. First, it can be noted that only one UE is scheduled in each time slot in the TDMA system. Let $\mathbf{H}_{r,k,t}$ denote the channel between RIS and UE k in time slot t. The corresponding transmit beamforming vector, the receive beamforming vector, and the RIS phase shift matrix are represented as $\mathbf{w}_{k,t}$, $\mathbf{f}_{k,t}$, and $\mathbf{\Phi}_{k,t}$, respectively. The sum rate can be accordingly rewritten as

$$\sum_{t=1}^{T} R_t = \sum_{t=1}^{T}\sum_{k=1}^{K} u_{k,t}\log_2\left(1+\frac{\left|\mathbf{f}_{k,t}^H(\mathbf{G}_t\mathbf{\Phi}_{k,t}\mathbf{H}_{r,k,t})^H\mathbf{w}_{k,t}\sqrt{P_T}\right|^2}{\sigma^2}\right) \tag{18}$$

Then, we assume that the channels are quasi-static and do not change over T time slots, so we have $\mathbf{G}_t = \mathbf{G}, \mathbf{H}_{r,k,t} = \mathbf{H}_{r,k},\ \forall t \in \{1,2,\ldots,T\}$. In this case, the beamforming vectors and the RIS phase shifts for UE k can be simplified to be consistent in different time slots, which can be represented as $\mathbf{w}_{k,t} = \mathbf{w}_k$, $\mathbf{f}_{k,t} = \mathbf{f}_k$, and $\mathbf{\Phi}_{k,t} = \mathbf{\Phi}_k,\ \forall t \in \{1,2,\ldots,T\}$. Accordingly, the sum rate can be further rewritten as

$$\sum_{t=1}^{T} R_t = \sum_{t=1}^{T}\sum_{k=1}^{K} u_{k,t}\log_2\left(1+\frac{\left|\mathbf{f}_k^H(\mathbf{G}\mathbf{\Phi}_k\mathbf{H}_{r,k})^H\mathbf{w}_k\sqrt{P_T}\right|^2}{\sigma^2}\right) \tag{19}$$

Let $R_k = \log_2(1+\frac{|\mathbf{f}_k^H(\mathbf{G}\mathbf{\Phi}_k\mathbf{H}_{r,k})^H\mathbf{w}_k\sqrt{P_T}|^2}{\sigma^2})$ represent the transmission rate of UE k. Thus, P1 can be rewritten as

$$\max_{\mathbf{U};\mathbf{w}_k,\mathbf{\Phi}_k,\mathbf{f}_k,\forall k} \sum_{t=1}^{T}\sum_{k=1}^{K} u_{k,t}R_k \tag{20}$$

$$\text{s.t.} \sum_{k=1}^{K} u_{k,t} = 1,\ \forall t \in \{1,2,\ldots,T\}, \tag{21}$$

$$\mathcal{A}_k \le \mathcal{A}_{k,\max},\ \forall k \in \{1,2,\ldots,K\}, \tag{22}$$

$$u_{k,t} \in \{0,1\}, \forall k \in \{1,2,\ldots,K\}, t \in \{1,2,\ldots,T\}, \tag{23}$$

$$\varphi_{m,k} \in \mathcal{F},\ \forall m \in \{1,2,\ldots,M\}, \forall k \in \{1,2,\ldots,K\}, \tag{24}$$

$$\mathbf{w}_k \in \Gamma_t, \forall k \in \{1,2,\ldots,K\}, \tag{25}$$

$$\mathbf{f}_k \in \Gamma_r, \forall k \in \{1,2,\ldots,K\}. \tag{26}$$

Note that the sum rate can also be converted as $\sum_{t=1}^{T}\sum_{k=1}^{K} u_{k,t}R_k = \sum_{k=1}^{K}(\sum_{t=1}^{T} u_{k,t})R_k$, where $\sum_{t=1}^{T} u_{k,t} \ge 0$, $\forall k \in \{1,2,\ldots,K\}$. Since the transmission rate R_k is related only to $\mathbf{w}_k$, $\mathbf{\Phi}_k$ and $\mathbf{f}_k$ and not to $\mathbf{U}$, and the transmission rates of different UEs are independent of each other, we can decompose P1 into K *per-UE rate maximization problems* and a *scheduling strategy design problem*.

1) Per-UE Rate Maximization Problem: This subproblem aims to maximize the achievable transmission rate of each UE through the joint optimization of beamforming vectors and the reflection-coefficient matrix of the RIS. We denote the achievable transmission rate of UE k as R_k, and the subproblem can be written as

$$\max_{\mathbf{w}_k,\mathbf{\Phi}_k,\mathbf{f}_k} R_k = \log_2\left(1+\frac{\left|{\mathbf{f}_k}^H{\mathbf{H}_k}^H\mathbf{w}_k\sqrt{P_T}\right|^2}{\sigma^2}\right) \tag{27}$$

$$\text{s.t.}\ \ \varphi_{m,k} \in \mathcal{F},\ \forall m \in \{1,2,\ldots,M\}, \tag{28}$$

$$\mathbf{w}_k \in \Gamma_t, \tag{29}$$

$$\mathbf{f}_k \in \Gamma_r. \tag{30}$$

In this subproblem, variables $\mathbf{w}_k$, $\mathbf{\Phi}_k$, and $\mathbf{f}_k$ are still coupled, so we further decompose the subproblem into a *beamforming optimization subproblem* and a *RIS reflection coefficient optimization subproblem*.

For the *beamforming optimization subproblem*, we assume that the reflection coefficient matrix of the RIS $\mathbf{\Phi}_k$ is fixed and maximize the transmission rate of each UE by optimizing the beamforming vectors $\mathbf{w}_k$ and $\mathbf{f}_k$. We can write this subproblem as

$$\max_{\mathbf{w}_k,\mathbf{f}_k} R_k = \log_2\left(1+\frac{\left|{\mathbf{f}_k}^H{\mathbf{H}_k}^H\mathbf{w}_k\sqrt{P_T}\right|^2}{\sigma^2}\right) \tag{31}$$

$$\text{s.t.}\ \ \mathbf{w}_k \in \Gamma_t, \tag{32}$$

$$\mathbf{f}_k \in \Gamma_r. \tag{33}$$

For the *RIS reflection coefficient optimization subproblem*, we fix the beamforming vectors and find the efficient reflection coefficient matrix $\mathbf{\Phi}_k$ for the RIS. We can write this subproblem as

$$\max_{\mathbf{\Phi}_k} R_k = \log_2\left(1+\frac{\left|{\mathbf{f}_k}^H{\mathbf{H}_k}^H\mathbf{w}_k\sqrt{P_T}\right|^2}{\sigma^2}\right) \tag{34}$$

$$\text{s.t.}\ \ \varphi_{m,k} \in \mathcal{F},\ \forall m \in \{1,2,\ldots,M\}. \tag{35}$$

2) Scheduling Strategy Design Problem: Based on the maximum achievable transmission rate of each UE, this subproblem is to design the scheduling strategy to maximize the total transmission rate over T time slots. The subproblem can be written as

$$\max_{\mathbf{U}} \sum_{t=1}^{T}\sum_{k=1}^{K} u_{k,t}R_k \tag{36}$$

$$\text{s.t.} \sum_{k=1}^{K} u_{k,t} = 1,\ \forall t \in \{1,2,\ldots,T\}, \tag{37}$$

$$\mathcal{A}_k \le \mathcal{A}_{k,\max},\ \forall k \in \{1,2,\ldots,K\}, \tag{38}$$

$$u_{k,t} \in \{0,1\}, \forall k \in \{1,2,\ldots,K\}, t \in \{1,2,\ldots,T\}. \tag{39}$$

In the next section, we will develop effective algorithms to solve the two decomposed problems and achieve the goal of sum rate maximization.

V. Sum Rate Maximization

In this section, we aim to propose a low-complexity algorithm to solve P1. Based on the decomposition of P1, the proposed solution consists of three parts: First, in Section V-A, we design a block coordinate descent (BCD)-based algorithm to solve the *per-UE rate maximization problem*, which solves the *beamforming optimization subproblem* and the *RIS reflection coefficient optimization subproblem* iteratively until the algorithm converges. Then, in Section V-B, we propose a heuristic scheduling algorithm to solve the *scheduling strategy design problem*. Finally, in Section V-C, we show the overall sum rate maximization algorithm for solving P1. The convergence and complexity analyses are given in Section V-D.

A. Per-UE Rate Maximization

To solve the *per-UE rate maximization problem*, we first design algorithms to solve the *beamforming optimization subproblem* and the *RIS reflection coefficient optimization subproblem*. Then we use the BCD algorithm to obtain the overall suboptimal solution. It is worth noting that we consider the difficulty of channel estimation in the RIS-assisted mmWave MIMO system. Thus, different from most existing studies adopting the BCD-based method [38], [40], we design the BCD algorithm in the case of unknown CSI.

1) Beamforming Optimization: Given the codebook of the BS and UE, although exhaustively searching all the transmit-receive beam pairs in the codebooks can find an efficient beam pair, we choose the hierarchical search method for reduced complexity. Specifically, we first design multilevel codebooks with different beam widths and then perform a divide-and-conquer search on the different codebook levels. The hierarchical search shows a tree structure, thereby substantially enhancing the search efficiency. The details of this method are given by Algorithm 1.

First, we focus on the design of the hierarchical codebooks Γ_t and Γ_r. There are two criteria to design a hierarchical codebook [41], which are given as follows.

- Within each layer, the aggregate beam coverage of all codewords should span the entirety of the angular domain, which ensures no miss of any angle during the beam search.
- The beam coverage of an arbitrary codeword within a layer should be given by the union of those of several adjacent codewords in the next layer, which establishes a tree-fashion relationship between the codewords.

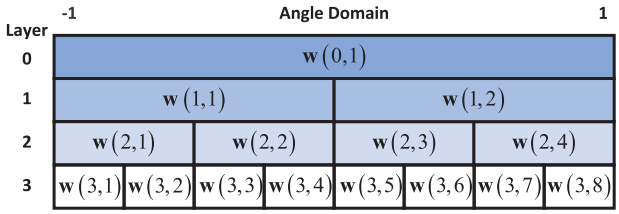

Fig. 3. Beam coverage of a 3-layer codebook.

In this article, we assume that each parent codeword has 2 child codewords, thus forming a binary-tree codebook structure. Fig. 3 shows a three-layer binary-tree codebook structure diagram, where $\mathbf{w}(l,n)$ denotes the n-th codeword of the l-th layer codebook. For antenna arrays with N antenna elements, we assume that there are N codewords covering the angle range [-1,1] in the last layer and each codeword has beam width $2/N$ with different steering angles. Therefore, the codebook consists of $\log_2(N)+1$ layers, where the k-th layer consists of 2^k codewords with beam width $2/2^k$ for each codeword.

For the design of N codewords in the last layer, since the steering vector $\mathbf{a}(N,\psi)$ in (7) can be defined as having a $2/N$ beam width centered on the steering angle ψ, we adopt the steering vectors with N angles evenly sampled within [-1,1] [41]. The n-th codeword exhibits the maximal beam gain along the angle $-1+\frac{2n-1}{N}$. Thus, we have $\mathbf{w}(\log_2(N),n) = \mathbf{a}(N,-1+\frac{2n-1}{N}),\ n=1,2,\ldots,N$. Then, for the design of codewords in the other layers, we use the joint sub-array and deactivation approach [41]. Specifically, to broaden the beam, we divide the N-antenna array into Q sub-arrays. Each sub-array is equipped with N_S antennas. Taking the first codeword of each layer (i.e., $\mathbf{w}(l,1)$) as an example, the codeword of the q-th sub-array can be represented as $\mathbf{w}_q = [\mathbf{w}(l,1)]_{(q-1)N_S+1:qN_S}$. Among these sub-arrays, the number of the activated sub-arrays is denoted by N_A. Since the beams of these activated sub-arrays are pointed in sufficiently spaced directions, they can be aggregated into wider beams. We define the codeword of the q-th activated sub-array as $\mathbf{w}_q = e^{j\theta_q}\mathbf{a}(N_S,-1+\frac{2q-1}{N_S})$ with $e^{j\theta_q}$ representing a scalar coefficient with the unit norm for the q-th sub-array. To reduce beam fluctuations, the intersection points between each sub-array coverage area are required to have high beam gain, which is modeled as the problem (27) in [41]. Based on the solution for the problem, we have $\theta_q = -q\frac{N_S-1}{N_S}\pi$. For the deactivated sub-arrays, the antennas in these sub-arrays are turned off, i.e., $\mathbf{w}_q = \mathbf{0}_{N_S\times 1}$. Thus, the codeword of each sub-array can be given by

$$\mathbf{w}_q = \begin{cases} e^{-jq\frac{N_S-1}{N_S}\pi}\mathbf{a}\left(N_S,-1+\frac{2q-1}{N_S}\right),\ q=1,2,\ldots,N_A \\ \mathbf{0}_{N_S\times 1},\ q=N_A+1,N_A+2,\ldots,Q, \end{cases} \tag{40}$$

and the beam width of the sub-array codeword is $\frac{2N_A}{N_S}$. In addition, according to Corollary 1 in [41], after obtaining the first codeword of each layer (i.e., $\mathbf{w}(l,1)$), we can obtain all the other codewords in the same layer through rotating $\mathbf{w}(l,1)$ by $\frac{2(n-1)}{2^l}$, $n=2,3,\ldots,2^l$, respectively. The beam rotation can be

Algorithm 1: Hierarchical Search Method for Beamforming Optimization.

Input: N_t; N_r; the designed hierarchical codebooks Γ_t and Γ_r
Output: $\mathbf{w}_k$; $\mathbf{f}_k$
Initialization: $cw_t = cw_r = 0$
1: Fix the BS to be in an omni-directional mode;
2: **for**each layer l_r in Γ_r**do**
3: Compare the data rate of the $(2cw_r - 1)$-th codeword with that of the $(2cw_r)$-th codeword and record the index of the codeword with a higher data rate as cw_r^*;
4: $cw_r = cw_r^*$;
5: **end for**
6: $\mathbf{f}_k = \Gamma_r(L_r, cw_r)$;
7: Fix the UE to the directional mode with $\mathbf{f}_k$;
8: **for**each layer l_t in Γ_t**do**
9: Compare the data rate of the $(2cw_t - 1)$-th codeword with that of the $(2cw_t)$-th codeword and record the index of the codeword with a higher data rate as cw_t^*;
10: $cw_t = cw_t^*$;
11: **end for**
12: $\mathbf{w}_k = \Gamma_t(L_t, cw_t)$;

realized by

$$\mathbf{w}(l,n) = \mathbf{w}(l,1) \circ \sqrt{N}\mathbf{a}\left(N, \frac{2(n-1)}{2^l}\right), n = 2,3,\ldots,2^l \tag{41}$$

where $\circ$ represents entry-wise product.

The details of the codebook design are presented in Algorithm 2. First, for the last layer of the codebook, the steering vectors with N angles evenly sampled within [-1,1] are used for codewords as in lines 2-3. Next, the joint sub-array and deactivation approach is adopted to generate the codewords for other layers as in lines 5-14. We first separate $\mathbf{w}(l,1)$ into $Q = 2^{\lfloor (p+1)/2 \rfloor}$ sub-arrays with $p = \log_2(N) - l$ in lines 5-6 and determine whether to activate half or all of the sub-arrays based on the parity of p in lines 7-11. Then, in line 12, the codebook for each sub-array can be obtained by (40) and we can get $\mathbf{w}(l,1)$ accordingly. In the end, based on $\mathbf{w}(l,1)$, we can derive all the other codewords in each layer by (41) in line 13. The hierarchical codebook design is finished after normalizing $\mathbf{w}(l,n)$ in line 14.

After obtaining the codebooks Γ_t and Γ_r, we first fix the BS in the omnidirectional mode and perform a binary tree search in Γ_r to find the efficiently received codeword for the UE, as in lines 1-6 in Algorithm 1. Specifically, in each layer, we select the codeword with a higher data rate, and the two adjacent codewords in the next-layer codebook within the beam coverage of this codeword are used as candidate codewords for the choice of the next layer. Then we fix the UE in the directional mode corresponding to the codeword and perform the same binary tree search in Γ_t to find the efficient transmit codeword for the BS in lines 7-12.

2) RIS Reflection Coefficient Optimization: For the RIS reflection coefficient optimization subproblem, we need to select

Algorithm 2: Hierarchical Codebook Design.

1: **for** each layer l**do**
2: **if** $l = \log_2(N)$ **then**
3: $\mathbf{w}(l,n) = \mathbf{a}(N, -1 + \frac{2n-1}{N}),\ n = 1,2,\ldots,N$;
4: **else**
5: $p = \log_2(N) - l$;
6: Separate $\mathbf{w}(l,1)$ into $Q = 2^{\lfloor (p+1)/2 \rfloor}$ sub-arrays with $\mathbf{w}_q = [\mathbf{w}(l,1)]_{(q-1)N_S+1:qN_S},\ q = 1,2,\ldots,Q$;
7: **if** p is odd **then**
8: $N_A = Q/2$;
9: **else**
10: $N_A = Q$;
11: **end if**
12: Calculate $\mathbf{w}_q$ according to (40) for $q = 1,2,\ldots,Q$ and obtain $\mathbf{w}(l,1)$;
13: Obtain all the other codewords in layer l through (41);
14: Normalize $\mathbf{w}(l,n)$;
15: **end if**
16: **end for**

Algorithm 3: Local Search Method for RIS Reflection Coefficient Optimization.

Input: M; b
Output: $\mathbf{\Phi}_k$
1: **for** $m = 1 : M$ **do**
2: $R_k^* = 0$;
3: **for** $p_s = 1 : 2^b$ **do**
4: Update $\mathbf{\Phi}_k$ with $\varphi_{m,k} = (p_s - 1)\frac{2\pi}{2^b - 1}$;
5: Obtain the transmission rate R_k;
6: **if** $R_k > R_k^*$ **then**
7: $R_k^* = R_k$, $\varphi_{m,k}^* = \varphi_{m,k}$;
8: **end if**
9: **end for**
10: Update $\mathbf{\Phi}_k$ with $\varphi_{m,k}^*$;
11: **end for**

the appropriate phase shift for each RIS element from a finite set of discrete phase shifts. Considering the complexity, we will use the local search method to solve the subproblem as shown in Algorithm 3. Specifically, we optimize each RIS element successively while keeping the phase shifts of the remaining $M - 1$ elements fixed. For each element, we traverse all the possible phase shifts and select the phase shift giving the maximum UE transmission rate as the optimized phase shift for the element. Then we use it for the phase shift optimization of other RIS elements until all the phase shifts are optimized.

3) Joint Optimization for Per-UE Rate Maximization: In order to solve the per-UE rate maximization problem, we apply the BCD method to alternately optimize the beamforming vectors and the RIS phase shift matrix. Specifically, as Algorithm 4 shows, we randomly initialize the beamforming vectors and the RIS phase shift matrix in the beginning. In each iteration, we

Algorithm 4: BCD Method for Joint Optimization of Beamforming and RIS Phase Shift.

Input: M; b; N_t; N_r; the designed hierarchical codebooks Γ_t and Γ_r
Output: $\mathbf{\Phi}_k^*$; $\mathbf{w}_k^*$; $\mathbf{f}_k^*$; R_k^*; γ_k^*
Initialization: $\tau = 0$; $R_k^0 = R_{bf} = 0$; $\delta = 3 \times 10^{-3}$; randomly generate $\mathbf{\Phi}_k^0$, $\mathbf{w}_k^0$ and $\mathbf{f}_k^0$; $\mathbf{\Phi}_k^* = \mathbf{\Phi}_k^0$; $\mathbf{w}_k^* = \mathbf{w}_k^0$; $\mathbf{f}_k^* = \mathbf{f}_k^0$

1: **repeat**
2: Obtain $\mathbf{f}_k^{\tau+1}$ and $\mathbf{w}_k^{\tau+1}$ with fixed $\mathbf{\Phi}_k^\tau$ using Algorithm 1;
3: Obtain data rate R_{bf} with $\mathbf{\Phi}_k^\tau$, $\mathbf{f}_k^{\tau+1}$ and $\mathbf{w}_k^{\tau+1}$;
4: **if** $R_k^\tau > R_{bf}$ **then**
5: $\mathbf{f}_k^{\tau+1} = \mathbf{f}_k^\tau$, $\mathbf{w}_k^{\tau+1} = \mathbf{w}_k^\tau$;
6: **end if**
7: Obtain $\mathbf{\Phi}_k^{\tau+1}$ with fixed $\mathbf{f}_k^{\tau+1}$ and $\mathbf{w}_k^{\tau+1}$ using Algorithm 3;
8: Obtain $R_k^{\tau+1}$ and SNR $\gamma_k^{\tau+1}$ with $\mathbf{\Phi}_k^{\tau+1}$, $\mathbf{f}_k^{\tau+1}$ and $\mathbf{w}_k^{\tau+1}$;
9: Update $\tau = \tau + 1$;
10: **until** $|R_k^\tau - R_k^{\tau-1}|/R_k^{\tau-1} < \delta$
11: Update $R_k^* = R_k^\tau$, $\gamma_k^* = \gamma_k^\tau$, $\mathbf{\Phi}_k^* = \mathbf{\Phi}_k^\tau$, $\mathbf{w}_k^* = \mathbf{w}_k^\tau$, $\mathbf{f}_k^* = \mathbf{f}_k^\tau$.

first fix the RIS phase shift matrix to the last updated value and use Algorithm 1 to update beamforming vectors. If the data rate of UE k after the beamforming update is more than that before this update, the results of this update are retained; otherwise, the beamforming vectors are not updated. Then, we update the RIS phase shift matrix based on Algorithm 3 with updated beamforming vectors. If the ratio of the difference in the data rate between two consecutive iterations is less than a certain threshold, i.e., $|R_k^\tau - R_k^{\tau-1}|/R_k^{\tau-1} < \delta$, we consider the algorithm has converged.

B. Scheduling Strategy Design

1) Motivation and Main Idea: After obtaining the maximum achieved rate of each UE with Algorithm 4, we need to design a scheduling strategy $\mathbf{U}$ to solve the scheduling strategy design problem. The difficulty in solving this problem lies in how to maximize the system sum rate while meeting the information freshness requirement of each UE. It should be noted that only when the SNR exceeds the threshold for reliable demodulation, can the AoI be reduced. Scheduling the UE with SNR below the threshold will not contribute to the information freshness requirement satisfaction and the sum rate enhancement. Thus, we filter UEs based on SNR and only schedule UEs with SNR above the threshold.

To schedule these filtered UEs, we design two scheduling phases. First, in *Scheduling Phase I*, we wish to ensure that the information freshness requirement of each UE is satisfied. Since we assume each UE receives the same types of service from the BS, which typically means the same information freshness requirements, we adopt a uniform and fair scheduling strategy. We schedule each UE in turn in the descending order of data rate. Each of the K UEs is scheduled once every K timeslots. Within a limited time T, this uniform and fair scheduling strategy can ensure that the time slot interval between two adjacent scheduling time slots is consistent for each UE, the difference in the number of scheduling time slots between UE with maximum rate and UE with minimum rate is not greater than once, and the maximum AoI of each UE over T timeslots will not exceed the number of UEs K. Therefore, each UE has similar AoI performance. Besides, according to Proposition 1 in [30], this uniform and fair scheduling strategy achieves the lower bound of the average episodic AoI, which is defined as $\frac{1}{K}\sum_{k=1}^{K}\mathcal{A}_k$. Motivated by these facts, we use the fair scheduling strategy to obtain better information freshness guarantees.

Then, in *Scheduling Phase II*, we wish to enhance the system sum rate as much as possible based on the scheduling result in Phase I. The main idea is to schedule UEs with the highest data rate as many times as possible without violating the AoI constraint of other UEs. Specifically, we select the UE with the highest data rate as the target UE and traverse all time slots. In each time slot, we replace the scheduled UE with the target UE and test the AoI of the originally scheduled UE. Only if the AoI satisfies the constraint (38), is the current replacement adopted.

2) Heuristic Scheduling Algorithm: The pseudocode of the heuristic scheduling algorithm is presented in Algorithm 5. For ease of presentation, we use $\mathcal{R}$ to denote the set of maximum achievable data rates for UEs without violating SNR constraints. The set of UEs without violating SNR constraints is represented by $\mathcal{K}_u$. The mapping between UE k and its data rate R_k^* is denoted by $k = \mathfrak{K}(R_k^*)$ and K_t is used to represent the scheduled UE in time slot t. First, we allocate time slots to UEs that can reliably demodulate the received signal in descending order of their data rates in *Scheduling Phase I*, which corresponds to lines 1-8. Specifically, in each time slot, we schedule the UE with the highest data rate in set $\mathcal{R}_t$, as in lines 2-3. In line 4, the data rate of the scheduled UE is removed from $\mathcal{R}_t$. In lines 5-7, if $\mathcal{R}_t$ is an empty set, which means that all UEs have been scheduled for one round, reinitialize $\mathcal{R}_t$ to $\mathcal{R}$ for the next round of scheduling. Next, we adjust the scheduling strategy to maximize the system sum rate as much as possible in *Scheduling Phase II*, as in lines 9-20. Specifically, we denote the UE with the maximum achievable data rate as $k^{\max}$. For each timeslot t, if UE $k^{\max}$ is not scheduled, we replace the scheduled UE K_t to UE $k^{\max}$, as in lines 11 and 12. In lines 13-18, we calculate the AoI of UE K_t and determine whether the AoI constraint is satisfied. If yes, then the adjustment is applied; otherwise, it is not applied. After completing the adjustments of all the timeslots, we obtain the final scheduling strategy.

C. Sum Rate Maximization

Following the design of the algorithms for the decomposed problems, we propose the sum rate maximization algorithm to solve P1. The algorithm is given by Algorithm 6. First, we generate the hierarchical codebooks for BS and each UE by Algorithm 2. Then we use Algorithm 4 to obtain the maximum achievable rate and the optimized variables for each UE. As in lines 5-8, we check the SNR of each UE to find the UEs

Algorithm 5: Scheduling Strategy Design.

Input: $\mathcal{R}$; $\mathcal{K}_u$; T; $\mathcal{A}_{k,\max}, \forall k$
Output: $\mathbf{U}^*$
Initialization: $u_{k,t}=0, \forall k,t$; $K_t=0, \forall t$; $\mathcal{R}_1=\mathcal{R}$
1: **for** each time slot t **do**
2: $\bar{k}=\mathfrak{K}(\max(\mathcal{R}_t))$;
3: $u_{\bar{k},t}=1$, $K_t=\bar{k}$;
4: $\mathcal{R}_{t+1}=\mathcal{R}_t-\{R^*_{\bar{k}}\}$;
5: **if** $\mathcal{R}_{t+1}=\emptyset$ **then**
6: $\mathcal{R}_{t+1}=\mathcal{R}$;
7: **end if**
8: **end for**
9: Treat the UE in $\mathcal{K}_u$ with the highest achievable data rate as the target UE, which is denoted as $k^{\max}$.
10: **for** each time slot t **do**
11: **if** $u_{k^{\max},t}=0$ **then**
12: $u_{k^{\max},t}=1, u_{K_t,t}=0$;
13: Calculate the AoI of UE K_t;
14: **if** $\mathcal{A}_{K_t}>\mathcal{A}_{K_t,\max}$ **then**
15: $u_{k^{\max},t}=0, u_{K_t,t}=1$;
16: **else**
17: $K_t=k^{\max}$;
18: **end if**
19: **end if**
20: **end for**

Algorithm 6: Sum Rate Maximization.

Input: M; b; K; T; N_t; N_r; $\mathcal{A}_{k,\max}, \forall k$; γ_{th}
Output: $\mathbf{U}^*$; $\mathbf{W}^*$; $\mathbf{\Phi}^*$; $\mathbf{F}^*$
Initialization: $\mathcal{R}=\emptyset$; $\mathcal{K}_u=\emptyset$
1: Generate the hierarchical codebook Γ_t for BS using Algorithm 2;
2: **for**each UE k**do**
3: Generate the hierarchical codebook Γ_r for UE k using Algorithm 2;
4: Obtain $\mathbf{\Phi}^*_k$, $\mathbf{w}^*_k$, $\mathbf{f}^*_k$, R^*_k and γ^*_k using Algorithm 4;
5: **if** SNR $\gamma^*_k>\gamma_{th}$ **then**
6: $\mathcal{R}=\mathcal{R}\cup\{R^*_k\}$;
7: $\mathcal{K}_u=\mathcal{K}_u\cup\{k\}$;
8: **end if**
9: **end for**
10: Obtain $\mathbf{U}^*$ using Algorithm 5;
11: Obtain $\mathbf{W}^*$, $\mathbf{\Phi}^*$ and $\mathbf{F}^*$ based $\mathbf{U}^*$;

which can achieve successful demodulation. The set $\mathcal{R}$ and $\mathcal{K}_u$ are obtained as the input of Algorithm 5 to get the scheduling strategy. In the end, since we have obtained $\mathbf{\Phi}^*_k$, $\mathbf{w}^*_k$, $\mathbf{f}^*_k$ for each UE k, we can easily obtain $\mathbf{W}^*$, $\mathbf{\Phi}^*$ and $\mathbf{F}^*$ according to the scheduling strategy $\mathbf{U}^*$.

D. Algorithm Analysis

1) Convergence Analysis: In Algorithm 6, we can see that the number of iterations for the scheduling algorithm (i.e., Algorithm 5) is fixed, while the number of iterations required to obtain the maximum achievable data rate for each UE (i.e., Algorithm 4) is uncertain. Thus, we focus on the convergence of Algorithm 4.

First, Theorem 1 shows that the objective function of the original optimization problem is non-decreasing in Algorithm 4.

Theorem 1: In Algorithm 4, $R(\mathbf{\Phi}_k^{\tau+1},\mathbf{w}_k^{\tau+1},\mathbf{f}_k^{\tau+1})\geq R(\mathbf{\Phi}_k^{\tau},\mathbf{w}_k^{\tau},\mathbf{f}_k^{\tau})$.

Proof: In line 2 of Algorithm 4, with $\mathbf{\Phi}_k^{\tau}$ given in the τ-th iteration, the beamforming vectors are obtained by Algorithm 1. It is worth noting that for Algorithm 1, an early search stage with weak beamforming gains is likely to experience relatively low SNR. This may lead to a higher probability of failing to find the best beam pair in the early search phase, resulting in subsequent misalignment at higher levels [42]. Considering this, in lines 4-6 of Algorithm 4, we compare R_k^{τ} and R_{bf} and decide whether to adopt the results of the hierarchical search. Therefore, $R(\mathbf{\Phi}_k^{\tau},\mathbf{w}_k^{\tau+1},\mathbf{f}_k^{\tau+1})\geq R(\mathbf{\Phi}_k^{\tau},\mathbf{w}_k^{\tau},\mathbf{f}_k^{\tau})$. Then, in line 7 of Algorithm 4, $\mathbf{\Phi}_k$ is updated using Algorithm 3 with the beamforming vectors fixed. The local search algorithm aims at maximizing the sum rate and searches for better phase shift values for each RIS element on the basis of $\mathbf{\Phi}_k^{\tau}$. Therefore, the performance of $\mathbf{\Phi}_k^{\tau+1}$ output by Algorithm 4 is better than or equal to the performance of $\mathbf{\Phi}_k^{\tau}$, which can be expressed as $R(\mathbf{\Phi}_k^{\tau+1},\mathbf{w}_k^{\tau+1},\mathbf{f}_k^{\tau+1})\geq R(\mathbf{\Phi}_k^{\tau},\mathbf{w}_k^{\tau+1},\mathbf{f}_k^{\tau+1})$. Thus, $R(\mathbf{\Phi}_k^{\tau+1},\mathbf{w}_k^{\tau+1},\mathbf{f}_k^{\tau+1})\geq R(\mathbf{\Phi}_k^{\tau},\mathbf{w}_k^{\tau},\mathbf{f}_k^{\tau})$. This completes the proof.

In addition, the number of discrete phase shifts and codewords for transmit and receive beamforming are limited, and the scheduling parameter is 0-1 variables, which makes the problem of maximizing the sum rate bounded and the output solutions guaranteed. Therefore, we have completed the proof of the convergence of the sum rate maximization algorithm.

2) Complexity Analysis: Since Algorithm 6 contains two parallel parts: the per-UE rate maximization and the scheduling strategy design, we analyze their complexity separately. First, for the per-UE rate maximization in lines 1-9, the for loop in line 2 has K iterations. In line 4, the complexity of Algorithm 4 is not only related to the number of iterations for the BCD method, which can be represented as N_{outer} to achieve the convergence condition $|R_k^{\tau}-R_k^{\tau-1}|/R_k^{\tau-1}<\delta$, but also related to the complexity of the beamforming optimization subproblem and RIS reflection coefficient optimization subproblem. For the former, two codewords are searched for at each layer of codebooks at both BS and UE. The complexity of the hierarchical search method is $O(2\log_2 N_t+2\log_2 N_r)$. For the latter one, the local search algorithm selects the best one among 2^b phase shifts for each element while keeping the phase shifts of the remaining elements unchanged. Since the RIS contains M elements, the complexity of this part is $O(M*2^b)$. Therefore, we get the complexity of Algorithm 4 as $O(N_{outer}*(2\log_2 N_t+2\log_2 N_r+M*2^b))$, and the complexity of the per-UE rate optimization is $O(K*(N_{outer}*(2\log_2 N_t+2\log_2 N_r+M*2^b)))$. Then, The scheduling strategy design corresponds to Algorithm 5. Both the for loops in line 1 and line 10 have T iterations, and The two for loops are parallel. Therefore, the complexity of Algorithm 5 is $O(T)$. In summary, the complexity of the sum rate maximization algorithm is $O(\max(K*(N_{outer}*(2\log_2 N_t+$

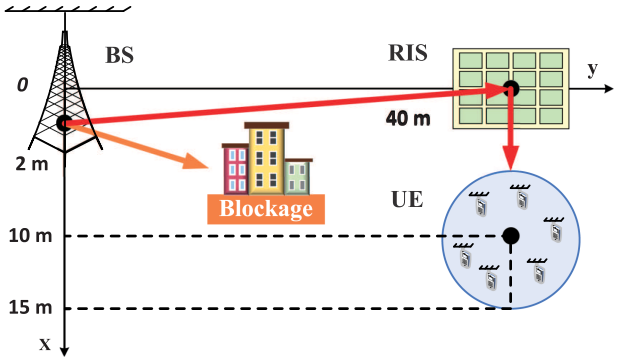

Fig. 4. Locations of communication nodes in the simulation.

$2\log_2 N_r + M * 2^b)), T))$. By simulation tests, N_{outer} ranges from 3 to 6. Such low complexity of the algorithm makes it suitable for practical implementation.

VI. Simulation Results and Discussions

In this section, we evaluate the performance of the proposed algorithm under various representative parameters. We also compare the performance of the proposed scheme with several baseline schemes and investigate the impact of different parameters on system performance.

A. Simulation Setup

In the simulation, we establish a Cartesian coordinate system to describe the locations of communication nodes. As shown in Fig. 4, the coordinates of the BS and the RIS are given by (2 m, 0 m) and (0 m, 40 m), respectively. UEs are uniformly distributed in a circle centered at (10 m, 40 m) with a radius of 5 m. The height of the BS, the RIS, and the UEs is set to 10 m, 2.5 m, and 1.5 m, respectively. The BS-RIS channel and the RIS-UE channel are generated according to the aforementioned SV model in LOS scenarios, which can be further written as

$$\mathbf{G} = \sqrt{\frac{N_t M}{P}}\left(\tilde{\alpha}_1 \mathbf{a}_r\left(M, \phi^r_{RIS,1}, \zeta^r_{RIS,1}\right)\mathbf{a}_t^H\left(N_t, \psi^t_{BS,1}\right) + \sum_{i=2}^{P}\tilde{\alpha}_i \mathbf{a}_r\left(M, \phi^r_{RIS,i}, \zeta^r_{RIS,i}\right)\mathbf{a}_t^H\left(N_t, \psi^t_{BS,i}\right)\right), \tag{42}$$

$$\mathbf{H}_{r,t} = \sqrt{\frac{MN_r}{L}}\left(\tilde{\beta}_1 \mathbf{a}_r\left(N_r, \psi^r_{UE,1}\right)\mathbf{a}_t^H\left(M, \phi^t_{RIS,1}, \zeta^t_{RIS,i}\right) + \sum_{i=2}^{L}\tilde{\beta}_i \mathbf{a}_r\left(N_r, \psi^r_{UE,i}\right)\mathbf{a}_t^H\left(M, \phi^t_{RIS,i}, \zeta^t_{RIS,i}\right)\right), \tag{43}$$

where $\tilde{\alpha}_1$ ($\tilde{\beta}_1$) $\sim \mathcal{CN}(0, 10^{-0.1\kappa})$ denotes the complex gain with the LOS component, $\tilde{\alpha}_i$ ($\tilde{\beta}_i$) $\sim \mathcal{CN}(0, 10^{-0.1(\kappa+\mu)})$ denotes the complex gain with the i-th NLOS path, and κ is the pathloss given by [43]

$$\kappa = a + 10b\log_{10}(\tilde{d}) + \xi, \tag{44}$$

TABLE I
Simulation Parameters

Parameter	Value
Transmit power P_T	45 dBm
Noise power σ^2	-90 dBm
Carrier frequency f_c	28 GHz
Termination iteration threshold δ	3×10^{-3}
SNR threshold value γ_{th}	2 dB
Number of path for BS-RIS channel P	4
Number of path for RIS-UE channel L	4

in which $\tilde{d}$ is the distance between the transmitter and receiver, and $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$. The values of a, b and σ_ξ are set as $a = 61.4$, $b = 2$, and $\sigma_\xi = 5.8$ dB as suggested by LOS real-world channel measurements [43]. The Rician factor μ is set to 10, which is defined as the ratio of the energy in the LOS path to the sum of the energy in other NLOS paths [13], [44]. In the following simulations, unless specified otherwise, we assume $K = 6$, $M_a = M_b = 10$, $N_t = N_r = 64$, $b = 3$, $T = 100$, and $\mathcal{A}_{k,\max} = \mathcal{A}_{\max} = 9, \forall k$. All simulation curves are averaged over 100 independent channel realizations. Other parameters are set as listed in Table I.

To validate the system performance of the proposed algorithm, we compare it with the following baseline algorithms:

1) *Random-RIS:* this algorithm randomly selects a feasible phase shift for each RIS element and keeps on using these phase shifts. Then, beamforming vectors are obtained by the hierarchical search method and the scheduling strategy is determined by Algorithm 5.
2) *Random-BF:* this algorithm randomly chooses the codewords from the codebooks for beamforming at both the BS side and the scheduled UE side. The codebook consists of all the code words in the last layer of the hierarchical codebook. Then, the RIS reflection coefficients are adjusted by the local research method and the scheduling strategy is computed by Algorithm 5.
3) *Round-Robin scheduling:* the only difference between this scheme and the proposed algorithm is the scheduling strategy. This scheme allocates time slots to UEs in descending order of data rates as in lines 1-7 of Algorithm 5, but it does not make further adjustments to the scheduling strategy.

B. Performance Evaluation

1) Impact of Maximum Tolerable AoI: In Figs. 5 and 6, we study the impact of the maximum tolerable AoI $\mathcal{A}_{\max}$ on the performance of the four schemes, which indicates the information freshness requirement of UEs. Specifically, Fig. 5 compares the sum rates of these schemes over T time slots under different $\mathcal{A}_{\max}$, and Fig. 6 compares the average system AoI of these schemes under different $\mathcal{A}_{\max}$, which is defined as the average AoI of all UEs, (i.e., $\frac{1}{K}\sum_{k=1}^{K}\mathcal{A}_k$). There are several important observations. First, the Round-Robin scheduling algorithm achieves the lowest average AoI, but the sum rate and the average AoI do not change with $\mathcal{A}_{\max}$. The reason is that it cannot adjust the time slot allocation according to the information freshness

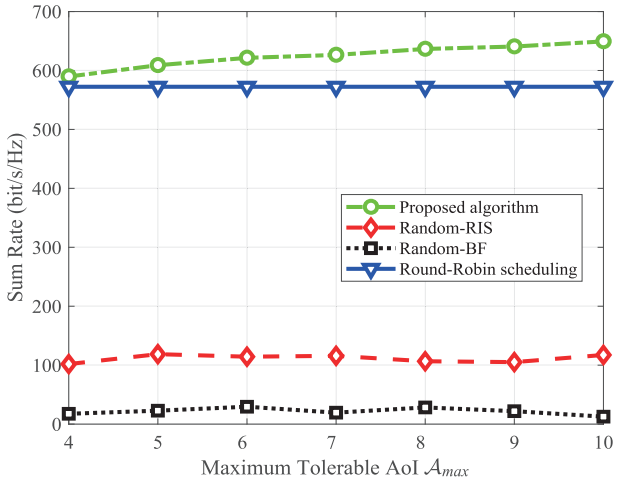

Fig. 5. Sum rate over T time slots versus $\mathcal{A}_{\max}$.

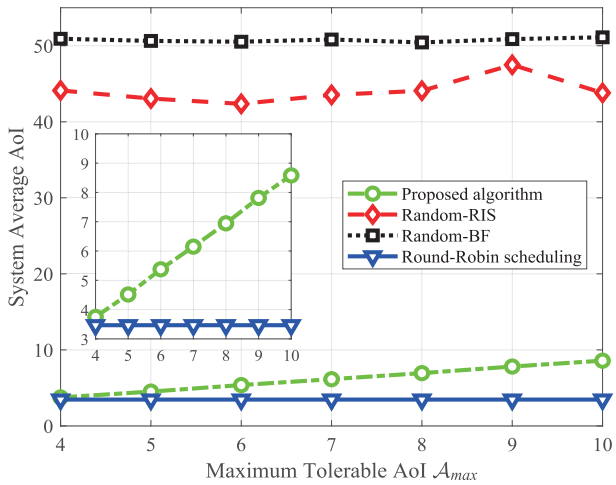

Fig. 6. System average AoI under different $\mathcal{A}_{\max}$.

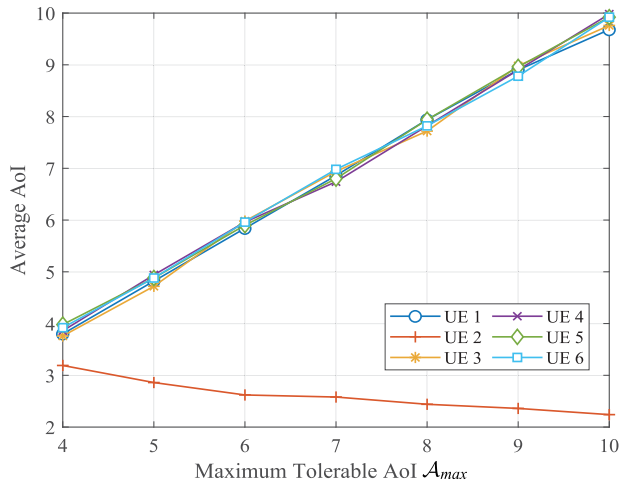

Fig. 7. Average AoI of each UE under different $\mathcal{A}_{\max}$.

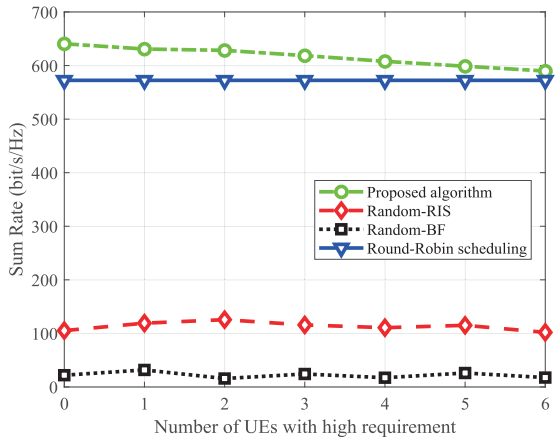

Fig. 8. Sum rate over T time slots versus the number of UEs with high requirement.

constraint. In other words, the scheduling strategy is consistent under different information freshness requirements, which limits the sum rate. In contrast, the proposed algorithm improves the sum rate performance at the cost of increasing the AoI while ensuring that the AoI constraints are satisfied. As $\mathcal{A}_{\max}$ is increased, the algorithm can increase AoI accordingly to obtain a larger sum rate. In addition, for the Random-BF scheme and Random-RIS scheme, both the data rates and the average AoI beyond $\mathcal{A}_{\max}$ are poor. This is because the random beamforming or random RIS reflection coefficients severely degrades the received signal quality, and even makes most UEs unable to demodulate the transmit signal. In this case, we treat the data rates of these UEs as zero. Accordingly, these UEs cannot be scheduled and the AoI of these UEs keeps on accumulating over time, which results in poor sum rate and AoI performance.

In Fig. 7, we focus on the proposed algorithm and show the average AoI performance of each UE under different $\mathcal{A}_{\max}$. Among all UEs, UE 2 has the highest data rate. First, we observe that the average AoI of each UE for different $\mathcal{A}_{\max}$ does not exceed $\mathcal{A}_{\max}$. Then, the larger the $\mathcal{A}_{\max}$, the smaller the average AoI of UE 2 and the larger the average AoI of other UEs. This is because as $\mathcal{A}_{\max}$ is increased, fewer time slots are needed to meet the information freshness requirements, so the proposed algorithm can allocate more time slots to UE 2 to enhance the sum rate over T time slots, which reduces the AoI of UE 2. This further explains the increases of the sum rate and average AoI of the proposed algorithm in Figs. 5 and 6 with increased $\mathcal{A}_{\max}$.

In Fig. 8, we consider the case where there are two optional service types on UEs, which correspond to different information freshness requirements. According to the service type of each UE, We divide UEs into two categories according to their information freshness requirements: UEs with high requirement and UEs with low requirement. The high-requirement corresponds to $\mathcal{A}_{k,\max} = 4$ and the low-requirement corresponds to $\mathcal{A}_{k,\max} = 9$. We plot the sum rate over T time slots while varying the number of UEs with high requirement from 0 to 6. When there are more high-requirement UEs in the system, the proposed algorithm needs to spend more time slots to satisfy the information freshness requirements, resulting in a lower sum rate. However, the proposed scheme still achieves the best performance among all the schemes.

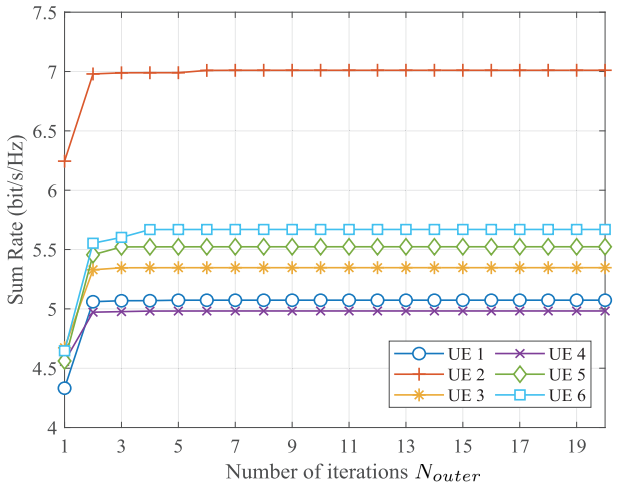

Fig. 9. Sum rate over T time slots versus number of iteration N_{outer}.

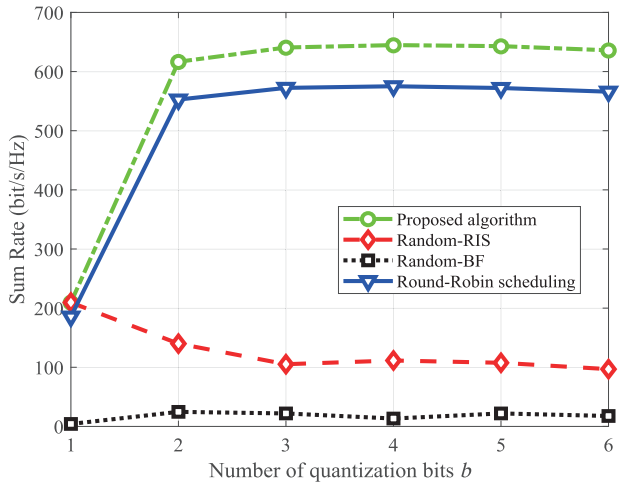

Fig. 11. Sum rate over T time slots versus the number of quantization bits b.

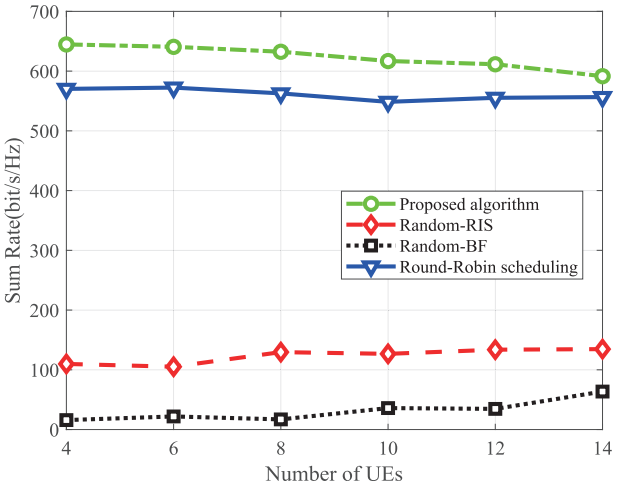

Fig. 10. Sum rate over T time slots versus the number of UEs.

2) Impact of Other Parameters: In Fig. 9, we examine the rate performance of each UE under the different number of iterations in Algorithm 4. We can see that the rate of all UEs converges to a stable value, which validates our convergence analysis in Section V-D. Besides, the number of iterations required for convergence is no more than 6, indicating that the BCD algorithm has a very fast convergence rate. Similar convergence rates can be seen in other related papers using the BCD algorithm, such as Figs. 3 and 4 in [15] and Fig. 3 in [45]. Such a fast convergence rate allows the algorithm to have reduced complexity.

In Fig. 10, we vary the number of UEs from 4 to 14 and compare the four schemes in terms of the sum rate over T time slots. Under the different number of UEs, we always set UE 2 as the UE with the highest rate. It is observed that the proposed algorithm achieves the highest sum rate. As the number of UEs is increased, the sum rate of the proposed algorithm shows a decreasing trend. This is because the scheduling strategy needs to meet the information freshness requirements for more UEs within T time slots, and the number of additional time slots allocated to the UE with the highest data rate is reduced accordingly. In contrast, the Round-Robin scheduling scheme does not take into account the information freshness of UEs, so the changes in the sum rate are only related to the rate performance of the added UEs. In general, when the number of UEs is 14, the performance gap between the proposed algorithm and Round-Robin scheduling is 6.28%. Further, a noticeable difference is observed between the proposed scheme and the other two schemes, i.e., Random-RIS and Random-BF, revealing the importance of jointly optimizing both RIS reflection coefficients and beamforming.

In Fig. 11, we plot the sum rates of the four schemes over T time slots while increasing the bit-quantization number from 1 to 6. As seen from the given results, the proposed algorithm outperforms the baseline schemes. The sum rates of the proposed scheme and the Round-Robin scheduling scheme gradually increase as b grows from 1 to 3, and then basically remain unchanged from 3 to 6. This shows that the system performance tends to be saturated when the number of quantization bits exceeds 3. The performance of the Random-RIS scheme is similar to that of the proposed scheme when b is 1. However, with the increase of b, it is more difficult to obtain an effective reflection coefficients matrix by Random-RIS. So the gap with the proposed algorithm widens when $b > 1$, and the sum rate fluctuates around a lower value. In addition, for Random-BF, when $b = 1$, the sum rate is close to 0, which means there are few UEs in the system which can reliably demodulate the transmit signal. As b increases, RIS can provide performance gain for reliable demodulation. However, due to the random beamforming, the beams between BS, RIS, and the scheduled UE are not well aligned, which impedes the growth of the sum rate.

In Fig. 12, we plot the sum rates over T time slots of the four schemes versus the number of RIS elements M. We see that the proposed algorithm outperforms the others as M is increased from 36 to 256. Then, all these four schemes show an increasing trend with M, which indicates that we can enhance the system sum rate by deploying RIS with more elements. Note that the increase of the Random-RIS scheme is due to the aperture gain of the RIS. The larger the RIS aperture, the more signal

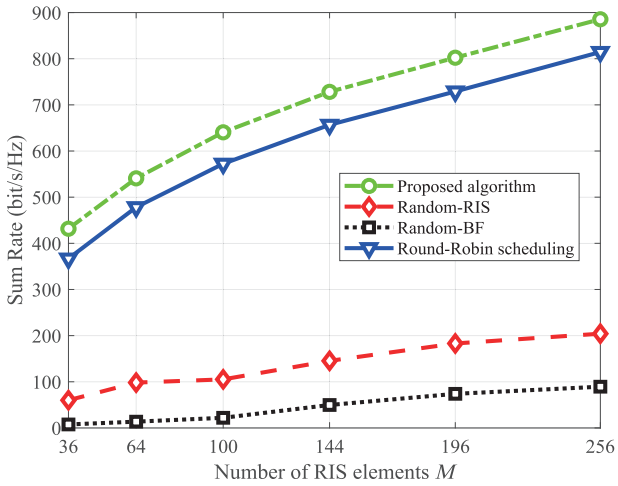

Fig. 12. Sum rate over T time slots versus the number of RIS elements M.

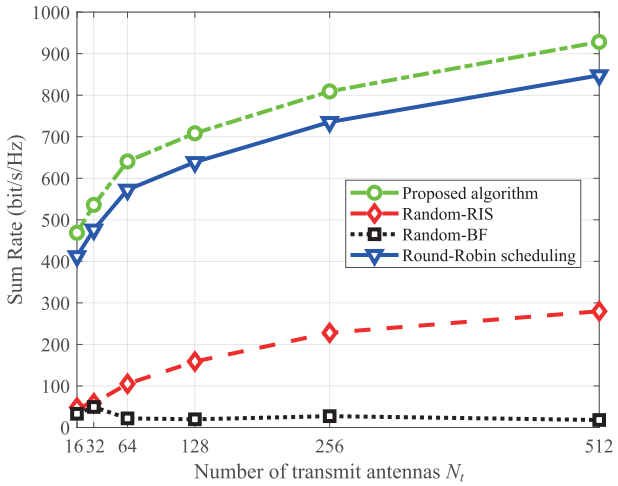

Fig. 13. Sum rate over T time slots versus the number of transmit antennas N_t.

power in the BS-RIS link can be collected by the RIS. Further, the gaps between the proposed algorithm and the other two schemes, i.e., Random-RIS and Random-BF, gradually widen as M is increased. Therefore, we need to design the joint RIS and beamforming optimization more carefully when more RIS elements are available.

In Fig. 13, we plot the sum rate over T time slots versus the number of transmit antenna N_t at the BS, which is varied from 16 to 512. It can be seen that the sum rate increases with the number of transmit antennas for the proposed algorithm, the Round-Robin scheduling scheme, and the Random-RIS scheme. The growths slow down with further increased number of antennas. When N_t is less than 128, the increase of N_t results in the most significant improvement in the sum rate. However, since the Random-BF scheme cannot provide a stable beamforming gain for the system, the increase in the number of transmit antennas has little impact on its performance. From this figure, we observe that the proposed algorithm has the highest sum rate than the other schemes. In general, when the number of transmit antennas

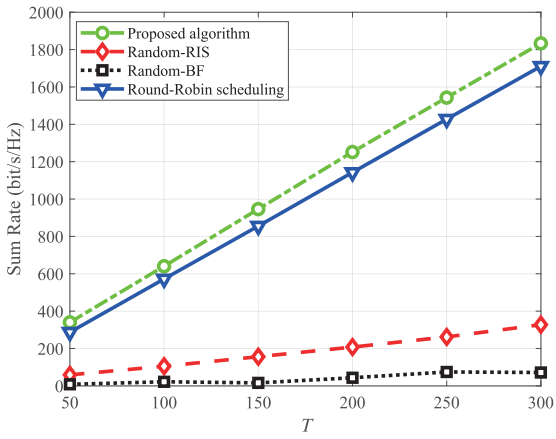

Fig. 14. Sum rate over T time slots versus T.

is 512, the performance gap between the proposed algorithm and the three baseline schemes is 9.5%, 231.7%, and 5039.3%, respectively.

In Fig. 14, we vary T from 50 to 300 and compare the four schemes in terms of the sum rate over T time slots. We can see that except for the Random-BF scheme, where the beams cannot be well aligned, the sum rates of all the other schemes increase linearly with T. The average sum rate of T time slots for the three schemes, i.e., $\frac{1}{T}\sum_{t=1}^{T} R_t$, can be calculated as 5.97 b/s/Hz, 5.69 b/s/Hz, and 1.07 b/s/Hz, respectively. Apparently, the proposed algorithm has the best sum rate performance. The reason is that the proposed scheme can schedule the UE with the highest data rate as many times as possible compared to Round-Robin scheduling, and it can achieve the joint optimization of beamforming and RIS reflection coefficients compared to Random-BF and Random-RIS.

VII. Conclusion

In this article, we investigated the sum rate maximization problem in RIS-assisted mmWave MIMO communication systems, where the information freshness requirements of all UEs should be satisfied. To solve this problem, we adopted the BCD method to jointly optimize RIS reflection coefficients and beamforming, and the heuristic scheduling algorithm to design the scheduling strategy. In particular, considering the difficulty of channel estimation in such systems, we utilized the hierarchical search method to update beamforming and the local search method to update RIS reflection coefficients. Simulation results showed that our algorithm can not only ensure the information freshness of UEs but also have the best sum rate performance. In future work, we will consider the case of scheduling multiple UEs in each time slot, where we will jointly design beamforming vectors, RIS phase shifts, and scheduling strategies to combat inter-user interference and satisfy the requirements of information freshness. In addition, we will extend this work to multi-cell multi-RIS scenarios in the future. With the joint design and optimization for multiple BSs and RISs, the information freshness requirement of UEs can be more effectively satisfied, and the system sum rate can be further improved.

REFERENCES

[1] IMT traffic estimates for the years 2020 to 2030, ITU, Geneva, Switzerland, Rep. M.2370-0 Jul. 2015. [Online]. Available: http://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2370-2015-PDF-E.pdf

[2] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[3] A. Ghosh et al., "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.

[4] S. Sun, T. S. Rappaport, M. Shafi, P. Tang, J. Zhang, and P. J. Smith, "Propagation models and performance evaluation for 5G millimeter-wave bands," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8422–8439, Sep. 2018.

[5] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of Information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[6] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, Nov. 2017.

[7] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heat, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[8] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Springer Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, Apr. 2015.

[9] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14960–14973, Dec. 2020.

[10] 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.

[11] P. Wang, J. Fang, W. Zhang, Z. Chen, H. Li, and W. Zhang, "Beam training and alignment for RIS-Assisted millimeter-wave systems: State of the art and beyond," *IEEE Wireless Commun.*, vol. 29, no. 6, pp. 64–71, Dec. 2022.

[12] N. S. Perovic, M. D. Renzo, and M. F. Flanagan, "Channel capacity optimization using reconfigurable intelligent surfaces in indoor mmWave environments," in *Proc. IEEE Int. Conf. Commun.*, 2020, pp. 1–7.

[13] P. Wang, J. Fang, L. Dai, and H. Li, "Joint transceiver and large intelligent surface design for massive MIMO mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1052–1064, Feb. 2021.

[14] C. Feng, W. Shen, J. An, and L. Hanzo, "Joint hybrid and passive RIS-Assisted beamforming for mmWave MIMO systems relying on dynamically configured subarrays," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13913–13926, Aug. 2022.

[15] R. Li, B. Guo, M. Tao, Y. -F. Liu, and W. Yu, "Joint design of hybrid beamforming and reflection coefficients in RIS-Aided mmWave MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2404–2416, Apr. 2022.

[16] P. Wang, J. Fang, W. Zhang, and H. Li, "Fast beam training and alignment for IRS-assisted millimeter wave/terahertz systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2710–2724, Apr. 2022.

[17] X. Wei, L. Dai, Y. Zhao, G. Yu, and X. Duan, "Codebook design and beam training for extremely large-scale RIS: Far-field or near-field?," *China Commun.*, vol. 19, no. 6, pp. 193–204, Jun. 2022.

[18] W. Wang and W. Zhang, "Joint beam training and positioning for intelligent reflecting surfaces assisted millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6282–6297, Oct. 2021.

[19] Q. He, D. Yuan, and A. Ephremides, "Optimal link scheduling for age minimization in wireless systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5381–5394, Jul. 2018.

[20] I. Kadota, A. Sinha, and E. Modiano, "Scheduling algorithms for optimizing Age of Information in wireless networks with throughput constraints," *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1359–1372, Aug. 2019.

[21] Q. Liu, H. Zeng, and M. Chen, "Minimizing AoI with throughput requirements in multi-path network communication," *IEEE/ACM Trans. Netw.*, vol. 30, no. 3, pp. 1203–1216, Jun. 2022.

[22] R. V. Bhat, R. Vaze, and M. Motani, "Throughput maximization with an average Age of Information constraint in fading channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 481–494, Jan. 2021.

[23] F. Wu, H. Zhang, J. Wu, Z. Han, H. V. Poor, and L. Song, "UAV-to-Device underlay communications: Age of Information minimization by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4461–4475, Jul. 2021.

[24] T. D. P. Perera, D. N. K. Jayakody, I. Pitas, and S. Garg, "Age of Information in SWIPT-Enabled wireless communication system for 5 GB," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 162–167, Oct. 2020.

[25] A. Muhammad, I. Sorkhoh, M. Samir, D. Ebrahimi, and C. Assi, "Minimizing Age of Information in multiaccess-edge-computing-assisted IoT networks," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13052–13066, Aug. 2022.

[26] J. Lee, D. Niyato, Y. L. Guan, and D. I. Kim, "Learning to schedule joint radar-communication with deep multi-agent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 406–422, Jan. 2022.

[27] I. Sorkhoh, M. A. Arfaoui, M. Khabbaz, and C. Assi, "Optimizing information freshness in RIS-Assisted cooperative autonomous driving," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 1518–1523.

[28] A. Muhammad, M. Elhattab, M. Shokry, and C. Assi, "Leveraging reconfigurable intelligent surface to minimize Age of Information in wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 2525–2530.

[29] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing Age of Information through aerial reconfigurable intelligent surfaces: A. deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[30] X. Fan, M. Liu, Y. Chen, S. Sun, Z. Li, and X. Guo, "RIS-Assisted UAV for fresh data collection in 3D urban environments: A. deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 632–647, Jan. 2023.

[31] X. Feng, S. Fu, F. Fang, and F. R. Yu, "Optimizing Age of Information in RIS-Assisted NOMA networks: A. deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2100–2104, Oct. 2022.

[32] W. Lyu, Y. Xiu, J. Zhao, and Z. Zhang, "Optimizing the Age of Information in RIS-Aided SWIPT networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2615–2619, Feb. 2023.

[33] Z. Shi, H. Wang, Y. Fu, X. Ye, G. Yang, and S. Ma, "Outage performance and AoI minimization of HARQ-IR-RIS aided IoT networks," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1740–1754, Mar. 2023.

[34] 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, Mar. 2020.

[35] 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.

[36] 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.

[37] M. Gao, B. Ai, Y. Niu, Z. Han, and Z. Zhong, "IRS-Assisted high-speed train communications: Outage probability minimization with statistical CSI," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[38] Y. Chen et al., "Reconfigurable intelligent surface assisted device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2792–2804, May 2021.

[39] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[40] N. Huang, T. Wang, Y. Wu, Q. Wu, and T. Q. S. Quek, "Integrated sensing and communication assisted mobile edge computing: An energy-efficient design via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2085–2089, Oct. 2022.

[41] Z. Xiao, T. He, P. Xia, and X. -G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[42] C. Liu, M. Li, S. V. Hanly, I. B. Collings, and P. Whiting, "Millimeter wave beam alignment: Large deviations analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1619–1631, Jul. 2017.

[43] M. R. Akdeniz et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[44] M. K. Samimi, G. R. MacCartney, S. Sun, and T. S. Rappaport, "28 GHz millimeter-wave ultrawideband small-scale fading models in wireless channels," in *Proc. IEEE Veh. Technol. Conf. Spring*, 2016, pp. 1–6.

[45] H. Gao, K. Cui, C. Huang, and C. Yuen, "Robust beamforming for RIS-Assisted wireless communications with discrete phase shifts," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2619–2623, Dec. 2021.

Ziqi Guo was born in Shandong, China, in 2000. He received the B.E. degree in communication engineering in 2021 from Beijing Jiaotong University, Beijing, China, where he is currently working toward the M.S. degree with the State Key Laboratory of Advanced Rail Autonomous Operation. His research interests include mmWave wireless communications and reconfigurable intelligent surface.

Yong Niu (Senior Member, IEEE) received the B.E. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 2011, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. From 2014 to 2015, he was a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently an Associate Professor with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University. His research interests include networking and communications, including millimeter wave communications, device-to-device communication, mediumaccess control, and software-defined networks. He was a Technical Program Committee Member for IWCMC 2017, VTC 2018-Spring, IWCMC 2018, INFOCOM 2018, and ICC 2018. He was the Session Chair for IWCMC 2017. He was the recipient of the Ph.D. National Scholarship of China in 2015, the Outstanding Ph.D. Graduates and Outstanding Doctoral Thesis of Tsinghua University in 2016, the Outstanding Ph.D. Graduates of Beijing in 2016, and the Outstanding Doctorate Dissertation Award from the Chinese Institute of Electronics in 2017, and the 2018 International Union of Radio Science Young Scientist Award.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. He is currently a Professor and Earle C. Williams Eminent Scholar Chair in electrical and computer engineering with Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid. He is a Distinguished Lecturer of IEEE Communications Society during 2021–2022, and IEEE Council of RFID during 2021–2022, and a Distinguished Lecturer during 2014–2018, and a Distinguished Speaker of IEEE Vehicular Technology Society during 2018–2021. He is on the Editorial Board of IEEE/CIC China Communications, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, ACM GetMobile, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE MULTIMEDIA, IEEE NETWORK, and IEEE NETWORKING LETTERS. He was the co-recipient of the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, IEEE ComSoc MMTC 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award, Best Demo Award of IEEE SECON 2017, Best Paper Awards of IEEE GLOBECOM 2019, 2016, and 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He isa Member of the ACM.

Changming Zhang received the B.S. degree from the Department of Electronic Information Science and Technology, Beijing Normal University, Beijing, China, in 2010, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, in 2015. He is currently a Research Expert with the Research Institute of Intelligent Networks, Zhejiang Lab, Hangzhou, China. His research interests include millimeter-wave and terahertz wireless communications, including huge-capacity transmission, complex digital signal processing, and broadband wireless networks.

Ning Wang (Member, IEEE) received the B.E. degree in communication engineering from Tianjin University, Tianjin, China, in 2004, the M.A.Sc. degree in electrical engineering from The University of British Columbia, Vancouver, BC, Canada, in 2010, and the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2013. From 2004 to 2008, he was with China Information Technology Design and Consulting Institute, as a Mobile Communication System Engineer, specializing in planning and design of commercial mobile communication networks, network traffic analysis, and radio network optimization. From 2013 to 2015, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia. Since 2015, he has been with the School of Information Engineering, Zhengzhou University, Zhengzhou, China, where he is currently an Associate Professor. He also holds adjunct appointments with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, and the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. His research interests include resource allocation and security designs of future cellular networks, channel modeling for wireless communications, statistical signal processing, and cooperative wireless communications. He has served on the technical program committees of international conferences, including IEEE GLOBECOM, IEEE ICC, IEEE WCNC, and CyberC. He was on the Finalist of the Governor Generals Gold Medal for Outstanding Graduating Doctoral Student from the University of Victoria in 2013.

Zhangdui Zhong (Fellow, IEEE) received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1988, respectively. He is currently a Professor and an Advisor of Ph.D. students with Beijing Jiaotong University, where he is also the Chief Scientist of the State Key Laboratory of Advanced Rail Autonomous Operation. He is currently the Director of the Innovative Research Team, Ministry of Education, Beijing, and the Chief Scientist of the Ministry of Railways, Beijing. He is an Executive Council Member of the Radio Association of China, Beijing, and the Deputy Director of the Radio Association, Beijing. He has authored or coauthored seven books, five invention patents, and more than 200 scientific research papers in his research field, which include wireless communications for railways, control theory, and techniques for railways, and GSM-R systems. His research has been widely used in railway engineering, such as the Qinghai-Xizang railway, Datong-CQinhuangdao Heavy Haul railway, and many high-speed railway lines in China. He was the recipient of the Mao YiSheng Scientific Award of China, Zhan TianYou Railway Honorary Award of China, and Top ten Science/Technology Achievements Award of Chinese Universities.

Bo Ai (Fellow, IEEE) received the M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2002 and 2004, respectively. He was an Excellent Post-Doctoral Research Fellow with Tsinghua University, Beijing, China, in 2007. He was a Visiting Professor with the EE Department, Stanford University, Stanford, CA, US, in 2015. He is currently with Beijing Jiaotong University as a Full Professor and a Ph.D. Candidate Advisor. He is also the Deputy Director of the State Key Laboratory of Advanced Rail Autonomous Operation and the Deputy Director of the International Joint Research Center. He is one of the main responsible people for Beijing Urban rail operation control system International Science and Technology Cooperation Base and the Member of the Innovative Engineering based jointly granted by Chinese Ministry of Education and the State Administration of Foreign Experts Affairs. He has authored/coauthored eight books and authored or coauthored more than 300 academic research papers in his research area. He has hold 26 invention patents. He has been the research team leader for 26 national projects and has won some important scientific research prizes. He has been notified by Council of Canadian Academies (CCA), that based on Scopus database. He has been listed as one of the Top one authors in his field all over the world. He has also been Feature Interviewed by *Electronics Letters* (IET). His research interests include the research and applications of channel measurement and channel modeling, dedicated mobile communications for rail traffic systems. He was the recipient of some important scientific research prizes.