IEEE ComSoc IEEE Communications Society | IEEE Open Journal of the Communications Society

Received 4 March 2026; accepted 25 March 2026. Date of publication 2 April 2026; date of current version 7 April 2026.

Digital Object Identifier 10.1109/OJCOMS.2026.3680322

Dynamic Resource Allocation in IRS-Assisted Hybrid mmWave and THz Networks for eURLLC and FeMBB Services

MUDDASIR RAHIM AND GEORGES KADDOUM (Senior Member, IEEE)

Department of Electrical Engineering, École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, QC H3C 1K3, Canada

CORRESPONDING AUTHOR: M. RAHIM (muddasir.rahim.1@ens.etsmtl.ca)

ABSTRACT This paper examines the coexistence of further enhanced mobile broadband (FeMBB) and extreme ultra-reliable low-latency communication (eURLLC) services within an intelligent reconfigurable surface (IRS)-assisted hybrid millimeter wave (mmWave) and terahertz (THz) network. To avoid the obstruction of the line-of-sight (LOS) link, we propose a hybrid deployment scheme that combines terrestrial IRSs and aerial IRSs. Additionally, we introduce a sub-IRS concept, dividing the IRS into multiple sub-IRSs to serve multiple users. The coexistence of FeMBB and eURLLC services in the same network poses a significant resource allocation (RA) challenge. To address this, we propose a multi-objective optimization problem that jointly optimizes power, user, and service allocations. The goal is to maximize the FeMBB data rate and eURLLC reliability, which is a challenging NP-hard mixed-integer nonlinear programming problem. To solve this, we employ a weighted sum method to convert it into a single-objective optimization problem, decomposing it into FeMBB and eURLLC RA sub-problems. Specifically, we introduce a many-to-many matching game to allocate IRSs to FeMBB users. Simulation results demonstrate the superior performance of the proposed scheme over baseline methods. Additionally, the proposed algorithm's sum rate for FeMBB users closely approximates that of an exhaustive search method but with significantly lower computational complexity.

INDEX TERMS Further enhanced mobile broadband (FeMBB), intelligent reconfigurable surface (IRS), millimeter wave (mmWave), terahertz (THz), extreme ultra-reliable low latency communication (eURLLC).

I. INTRODUCTION

THE rapid growth of internet-of-everything (IoE) applications, with their diverse requirements, has significantly increased the demand for high data rates, low latency, and highly reliable transmissions. As a result, effective resource allocation (RA) has become more critical [1]. Fifth-generation (5G) networks currently support a variety of services, including massive machine-type communications (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable low-latency communications (URLLC) [2]. Additionally, sixth-generation (6G) communication networks will not only enhance these 5G services with ultra mMTC (umMTC), further eMBB (FeMBB), and extreme URLLC (eURLLC), but also support multiple services within the network [3].

To support emerging IoE applications, such as virtual reality (VR) and augmented reality (AR), it is necessary to ensure high data rates, ultra-reliability, and low latency. The vision of 6G is to enhance data rate performance, achieve extremely low latency, and enable ubiquitous connectivity [4]. This enhancement is envisioned through the coexistence of FeMBB and eURLLC services within the same network. However, managing resources in the presence of diverse service requirements poses a significant challenge [5], [6]. This challenge arises from the differing requirements of each service: FeMBB emphasizes transmitting large data blocks at high rates. At the same time, eURLLC focuses on delivering short packets with low latency and extremely high reliability. A recent concept known as resource puncturing has been proposed to facilitate the concurrent operation of eURLLC and FeMBB services within the same network [7]. In this approach, some resources initially allocated to FeMBB services are punctured and reallocated to eURLLC services [8], [9], [10]. FeMBB services demand a high data rate, while eURLLC services' performance is determined by reliability and latency based on the channel conditions among coexisting FeMBB and eURLLC user equipment (UE). Realizing these ambitious objectives

necessitates exploring innovative solutions that facilitate the coexistence of diverse services with varying requirements.

To address these challenges, millimeter wave (mmWave) and terahertz (THz) communications have emerged as promising technologies for future 6G wireless networks, offering ultra-wideband bandwidth and extremely high transmission rates [11], [12]. Despite their potential, there are significant challenges in their commercial deployments. Specifically, mmWave and THz bands suffer considerable signal attenuation and diffraction, which limit coverage range [11]. Solutions like ultra-massive multiple-input multiple-output (UM-MIMO) and intelligent reflecting surfaces (IRSs) can help overcome these limitations. However, UM-MIMO is associated with high energy consumption and hardware costs. Additionally, the high penetration and path loss inherent in mmWave and THz bands necessitate the use of ultra-dense pico-cells to cover blind spots, which can be costly, especially in environments with weak line-of-sight (LOS), such as dense urban areas.

Recently, IRSs have gained prominence for their ability to control wireless channels at a significantly lower cost. IRSs consist of a large array of compact passive elements and metamaterial-based structures designed to reflect incident signals [13]. These surfaces enhance spectrum utilization and reduce energy consumption in wireless communication systems. By reflecting incident signals, IRSs can alter the signal's phase, amplitude, and frequency. When direct communication experiences poor channel quality conditions, IRSs can improve transmissions between the transmitter and receiver. By leveraging IRSs, it is possible to enhance signal quality and extend coverage without the high expenses associated with traditional solutions like UM-MIMO, thereby addressing the key challenges faced by mmWave and THz communications [14]. Recently, advanced IRS technologies such as beyond-diagonal IRS (BD-IRS), stacked intelligent metasurfaces (SIMs), simultaneously transmitting and reflecting IRS (STAR-RIS), and learning-based IRS architectures have been investigated in cell-free networks [15], [16], [17]. Most IRS-assisted networks focus on the deployment of terrestrial IRS (TIRS) in scenarios without blockages [18], [19]. While TIRS can be easily installed on building facades, high blockage probabilities in dense urban environments pose challenges for covering blind spots flexibly. In contrast, aerial IRS (AIRS) schemes provide LOS links, offering a more effective solution for overcoming coverage limitations and enhancing network performance in obstructed environments. In this work, we propose a hybrid IRS (HIRS) deployment scheme that combines one static AIRS with multiple TIRSs to assist communication between the access point (AP) and UEs.

A. PRIOR WORKS

Considerable works have explored the allocation of URLLC traffic over eMBB traffic [20], [21], [22], [23], [24], [25]. In [20], the authors proposed a joint scheduling method aimed at achieving long-term quality of service (QoS) between eMBB and URLLC services. The proposed approach aimed to maximize the reliability of both services system-wide using multi-objective optimization problems (MOP). MOPs were simplified into single-objective optimization problems (SOP) using optimal Pareto mappings. Specifically, scheduling decisions utilized deep deterministic policy gradients (DDPG) with prioritized sampling, accounting for the unpredictability of URLLC traffic and channel variations. Similarly, in [21], the authors formulated a MOP focusing on the QoS of both eMBB and URLLC services. This work also proposed an improved Gale-Shapley algorithm for efficient traffic scheduling, leveraging the analytic hierarchy process method and coalitional game theory. In [22], the authors investigated an RA problem in ultra-dense networks (UDN) with coexisting eMBB and URLLC users. The proposed energy efficiency (EE) optimization problem with QoS constraints addresses diverse user requirements and interference issues. To solve the combinatorial non-convex fractional programming problem, this work decomposed it into resource block (RB) and power allocation subproblems, which were addressed using successive convex approximation (SCA) and difference of convex (DC) programming. An alternative optimization-based algorithm was proposed to optimize RB and power allocation jointly.

In [23], the authors integrated IRSs into cellular networks to enable the coexistence of URLLC and eMBB services. Their approach jointly optimized the power allocation at the base station (BS) and the IRS phase shift (PS) matrix to maximize the eMBB sum rate while satisfying the eMBB rate constraint. Similarly, [24] explored the use of IRSs to facilitate the coexistence of eMBB and URLLC services. This study formulated two optimization problems: one for allocating time slots for eMBB traffic and another for allocating mini-time slots for URLLC traffic, aiming to maximize the eMBB sum rate and the number of admitted URLLC packets while meeting both services' QoS requirements. To address computational complexity, they proposed proactively designed IRS phase-shift matrices optimized at the beginning of each time slot. Additionally, the study conducted in [25] addressed the RA problem in the coexistence of URLLC and eMBB services with IRS assistance in the THz band. This work formulated a joint optimization problem involving the transmit power, the reflection coefficient of the IRS elements, and the allocation of THz resource blocks. Their approach combined an optimization policy with deep learning and ensemble learning methods to meet the requirements of both eMBB and URLLC services.

Few research studies have investigated hybrid multi-band networks (MBNs) and HIRS deployments [26], [27], [28], [29], [30]. The authors of [31] explored the RA challenges in MBNs. The aim of this work was to provide a service to FeMBB and eURLLC in MBNs. The problem was formulated as a joint network and subchannel allocation problem, and a machine learning-based solution was proposed. In [28],

the authors introduced a dynamic AP clustering model to enable user-centric AP cooperation in a hybrid network composed of THz APs, mmWave APs, and sub-6GHz APs. In this user-centric model, each user initially formed three candidate clusters of APs, each from a different tier. One of these candidate clusters was then selected to serve the user based on either the maximum signal-to-interference-plus-noise ratio (SINR) cluster selection (MSCS) or the maximum rate cluster selection (MRCS) scheme. Stochastic geometry was utilized to establish an analytical framework for calculating the SINR and rate coverage probabilities of the hybrid user-centric networks.

In [29], the authors presented a hybrid transmission scheme designed to enhance the reliability of URLLC services in mmWave bands through adaptive diversity combining. This scheme integrated two links: one in the FR1 band, known for its higher reliability due to favorable propagation conditions, and another in the FR2 band, which provides greater capacity but less reliable connectivity. The hybrid scheme dynamically adjusts the use of both links to leverage the complementary characteristics of the FR1 and FR2 bands, switching between FR2-only and joint FR1-FR2 transmission based on the instantaneous channel quality of the primary FR2 link. Performance was assessed using selection combining (SC) and maximal ratio combining (MRC) techniques. In [30], the authors considered the hybrid deployment of IRSs as TIRSs and AIRS to avoid the obstruction of LOS communication in dense urban environments. The objective was to maximize the minimum average rate of users in different blind zones of dense urban areas by jointly optimizing transmit beamforming at the BS, the reflecting coefficients of TIRSs and AIRS, and the height of the deployed AIRS. This work proposed fractional programming (FP) for the beamforming at the BS, the projected subgradient method (PSM) for passive beamforming at the AIRS, and the closed-form expression of the TIRS reflection coefficient was derived.

B. CONTRIBUTIONS

The coexistence of FeMBB and eURLLC services within a single network presents a complex RA problem, as FeMBB and eURLLC services have different requirements, including the trade-offs between reliability, latency, and data rate. In addition, the unpredictable and sporadic nature of eURLLC traffic necessitates dynamic and intelligent RA strategies. This paper proposes a comprehensive RA framework that integrates IRS-assisted hybrid mmWave and THz communication with a matching theory-based approach. The proposed optimization framework incorporates delay constraints alongside minimizing the link error probability, ensuring compliance with the 3rd Generation Partnership Project (3GPP) defined eURLLC reliability requirements. This guarantees that packets are delivered within the specified latency deadlines and meet both link- and system-level performance metrics. This framework aims to effectively manage the coexistence of FeMBB and eURLLC services in the hybrid wireless network. The key contributions of this paper can be summarized as follows:

- We model the coexistence problem of FeMBB and eURLLC, representing FeMBB UEs' rate with the Shannon rate and using finite block length codes to characterize eURLLC UEs' rate in the IRS-assisted hybrid network. In addition, we introduce a hybrid deployment of TIRS and AIRS, which consists of multiple sub-IRSs.
- One of the main contributions of this paper is the formulation of an RA problem aimed at jointly optimizing key system parameters, such as the AP transmit power and UEs association. To maximize the data rate for FeMBB UEs and ensure reliability for eURLLC UEs, this optimization problem is framed as an MOP. The formulated problem is an NP-hard mixed-integer nonlinear programming (MINLP) problem, which is generally difficult to solve directly. To tackle the MOP, we employ the weighted sum method, transforming it into an SOP that still retains the MINLP characteristics.
- We decompose the SOP into two sub-problems: RA for FeMBB UEs and RA for eURLLC UEs. Each RA problem is further decomposed into UEs association and power allocation for both FeMBB and eURLLC UEs. For the FeMBB UEs association, we propose a Gale-Shapley-based many-to-many matching algorithm. Similarly, for the eURLLC UEs association, we propose a Gale-Shapley-based one-to-one matching algorithm. Furthermore, we derive a closed-form expression for the power allocation of both FeMBB and eURLLC UEs.
- Numerical results demonstrate that the FeMBB UEs' sum rate achieved by the proposed algorithm is comparable to that obtained using the exhaustive search (ES) method. In addition, as compared to baseline methods, the proposed algorithm significantly improves the sum rate for both FeMBB and eURLLC UEs and reliability for eURLLC UEs.

The rest of the paper is organized as follows. Section II illustrates the system model, channel modeling, data rate formulation, and mobility model. The RA optimization problem is formulated in Section III. The proposed solution is provided in Section IV. Furthermore, in Section V, the performance of the proposed approach is evaluated and compared with baseline schemes. Finally, the paper is concluded in Section VI.

Notations: Boldface lowercase symbols represent vectors, while boldface uppercase symbols represent matrices. Moreover, $\mathbf{X}^{\mathsf{T}}$ represents the transpose and $\mathbf{X}^{\mathsf{H}}$ represents the conjugate transpose (Hermitian) of the matrix $\mathbf{X}$. The absolute value and L2-norm operators are denoted as $|.|$ and $\|.\|$, respectively. The set $\mathbb{R}^{M\times 1}$ represents a real vector and $\mathbb{C}^{M\times 1}$ denotes a complex vector of dimensions $M \times 1$. Key network parameters and channel variables used in this paper are summarized in Table 1.

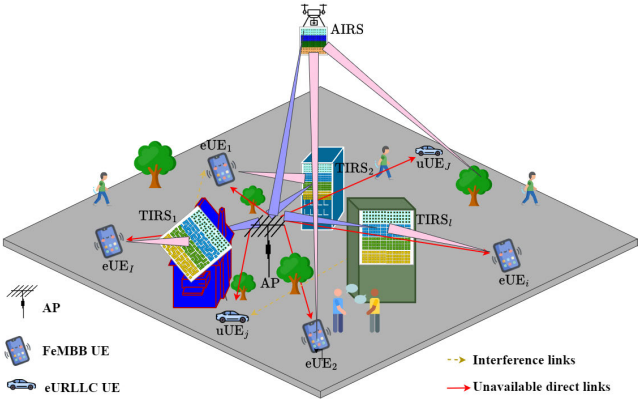


FIGURE 1. System model for coexisting FeMBB and eURLLC services in IRS-assisted multi-band networks.

II. SYSTEM MODEL

A. NETWORK SETTING

With reference to Fig. 1, we consider an IRS[1]-assisted network with one AP equipped with N antennas. The AP in the network provides services to K single-antenna UEs. Let us consider $J \in [1, K]$ as the number of eURLLC UEs in the network, then the number of FeMBB UEs is $I = K - J$. The sets of eURLLC and FeMBB UEs are denoted as $\mathcal{J} = \{\mathtt{uUE}_1, \mathtt{uUE}_2, \ldots, \mathtt{uUE}_j, \ldots, \mathtt{uUE}_J\}$ and $\mathcal{I} = \{\mathtt{eUE}_1, \mathtt{eUE}_2, \ldots, \mathtt{eUE}_i, \ldots, \mathtt{eUE}_I\}$, respectively. Furthermore, L IRSs are deployed in the network to assist the communication, where one AIRS and $L-1$ TIRSs are used. The set of L available IRSs is represented as $\mathcal{L} = \{\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{I}_l, \ldots, \mathrm{I}_L\}$. Furthermore, each IRS is equipped with $M = M_y M_z$ reflecting elements, where M_y and M_z denote the number of reflecting elements in each row and column of the IRS, respectively. The inter-element spacing along the y-axis and z-axis of the IRS elements is represented as Δ_y and Δ_z, respectively. Let us consider a multi-user scenario where the AP employs beamforming targeting-specific IRSs for corresponding UEs with optimal PS configuration at the IRS. We assume that all direct links between the AP and UEs are blocked due to obstacles[2]

B. SUB-IRS

We divide the IRS into $\mathcal{C}$ equal-sized sub-IRS, where each sub-IRS has $M_{\mathcal{C}} = M/\mathcal{C}$ reflecting elements. When an IRS panel is partitioned into multiple sub-IRSs, potential electromagnetic coupling or sidelobe leakage between adjacent sub-IRSs may arise in practical implementations. In this

[1]A conventional diagonal IRS model with ideal unit-modulus phase shifts and perfect synchronization is assumed. Mutual coupling, radiation loss, and other hardware impairments are not explicitly modeled in order to preserve analytical tractability and focus on the proposed sub-IRS resource allocation framework.

[2]This assumption is reasonable when the direct link is severely impeded by obstacles or human bodies in the environment [32], [33].. In addition, given the static nature of AP and IRSs LOS links between the AP and the IRSs can be established with a careful deployment of IRSs. However, due to the mobility of UEs, maintaining LOS links between all UEs and IRSs is impractical. Therefore, we assume that all AP-to-IRSs links have a LOS component available, while availability of the LOS depends on the location of the UE for the IRS-to-UE link.

TABLE 1. List of notations.

Notation	Description
I	Number of FeMBB UEs
J	Number of eURLLC UEs
K	Number of total UEs
L	Number of IRSs
M	Number of IRS reflecting elements
C	Number of sub-IRSs
M_c	Number of sub-IRS reflecting elements
$\mathtt{eUE}_i$	ith FeMBB UE
$\mathtt{uUE}_j$	jth eURLLC UE
I_l	lth IRS
$\mathrm{I}_{l,c}$	cth sub-IRS on I_l
$\mathrm{I}_{l,c,m}$	mth reflecting element on $\mathrm{I}_{l,c}$
$\mathbf{p}_a$	Center location of AP
$\mathbf{p}_{l,c}$	Center location of $\mathrm{I}_{l,c}$
$\mathbf{p}_k$	Center location of $\mathtt{UE}_k$
$r_{a,l,c}$	Distance between AP to $\mathrm{I}_{l,c}$
$r_{l,c,k}$	Distance between $\mathrm{I}_{l,c}$ to $\mathtt{UE}_k$
ψ	Evaluation angle
φ	Azimuth angle
$g^{\chi}_{n,(l,c,m)} \in \mathbb{C}^{M_c \times 1}$	Channel from $\mathtt{AP}_n$-to-$\mathrm{I}_{l,c,m}$
$g^{\chi}_{(l,c,m),k} \in \mathbb{C}^{M_c \times 1}$	Channel from $\mathrm{I}_{l,c,m}$-to-$\mathtt{UE}_k$
$\mathbf{G}^{\chi}_{l,c} \in \mathbb{C}^{M_c \times N}$	Channel matrix from AP-to-$\mathrm{I}_{l,c}$
$\mathbf{h}^{\chi}_{l,c,k}$	Cascaded channel
$\ell^{\chi}_{n,(l,c,m)}$	Pathloss in $\mathtt{AP}_n$-to-$\mathrm{I}_{l,c,m}$ link
$\ell^{\chi}_{(l,c,m),k)}$	Pathloss in $\mathrm{I}_{l,c,m}$ to $\mathtt{UE}_k$ link
ℓ^{χ}_k	Overall pathloss in the cascaded link
$\mathbf{\Theta}_{l,c}$	PS matrix of $\mathrm{I}_{l,c}$
$\kappa_{l,c,m} \in (0,1]$	Reflection coefficient of $\mathrm{I}_{l,c,m}$
$\theta_{l,c,m} \in [0, 2\pi)$	PS amplitude applied by $\mathrm{I}_{l,c,m}$
$\mathbf{w}_k$	Beamforming vector for $\mathtt{UE}_k$
$\mathbb{G}_{l,c,m}(-\mathbf{r}_k)$	Gains of $\mathrm{I}_{l,c,m}$ in reflection direction $\mathbf{r}_k$
$\mathbb{G}_{l,c,m}(\mathbf{r}_a)$	Gains of $\mathrm{I}_{l,c,m}$ in incident direction $\mathbf{r}_a$
$\mathbb{G}_a$	Antenna gain at AP
$\mathbb{G}_k$	Antenna gain at $\mathtt{UE}_k$
p_i	Transmit power of $\mathtt{eUE}_i$
p_j	Transmit power of $\mathtt{uUE}_j$
x_i	Message intended for $\mathtt{eUE}_i$
x_j	Message intended for $\mathtt{uUE}_j$
$y^u_{l,c,j}$	Received signal at $\mathtt{uUE}_j$ through $\mathrm{I}_{l,c}$
$y^e_{l,c,i}$	Received signal at $\mathtt{eUE}_i$ through $\mathrm{I}_{l,c}$
$\gamma^e_{l,c,i}$	SINR at $\mathtt{eUE}_i$ through $\mathrm{I}_{l,c}$
$\gamma^u_{l,c,j}$	SINR at $\mathtt{uUE}_j$ through $\mathrm{I}_{l,c}$
$R^e_{l,c,i}$	Data rate of $\mathtt{eUE}_i$ through $\mathrm{I}_{l,c}$
$R^u_{l,c,j}$	Data rate of $\mathtt{uUE}_j$ through $\mathrm{I}_{l,c}$
ζ_j	$\mathtt{uUE}_j$ packet length
V_j	$\mathtt{uUE}_j$ channel dispersion
ϵ_j	$\mathtt{uUE}_j$ error probability
τ_j	Delay requirement of $\mathtt{uUE}_j$

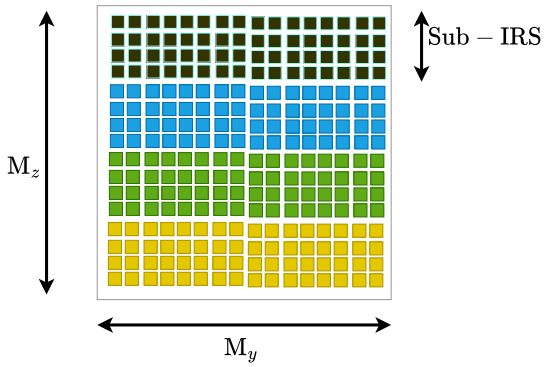


FIGURE 2. IRS with $M_y \times M_z$ reflecting elements is divided into $\mathcal{C}$ sub-IRSs each sub-IRS equipped with $M_{\mathcal{C}}$.

study, sub-IRSs on the same panel are assumed to be electrically isolated, and intra-IRS spillover effects are not explicitly modeled. This assumption is made to preserve analytical tractability and aligns with commonly used ideal IRS models that consider independent and controllable reflecting elements. While sidelobe interference could impact performance in dense or closely spaced sub-IRS configurations, modeling such effects would require a detailed electromagnetic characterization and is left for future investigation. We can perform simultaneous beamforming in all directions by dividing the IRS such that the dimension of the channel space for each sub-IRS equals the total number of sub-IRSs. We use the degrees-of-freedom (DOF) approximation to determine the number of elements in each sub-IRS, which can be written as [34]

$$\mathcal{C} = \frac{M}{M_{\mathcal{C}}} = \frac{\pi}{\lambda^2} M_{\mathcal{C}} \Delta_y \Delta_z, \tag{1}$$

where λ is the wavelength, which is evaluated at the center frequency of the operating bandwidth to determine the number of sub-IRSs. This choice reflects a practical design assumption, where the sub-IRS partitioning is based on dominant spatial characteristics and fixed hardware constraints. To calculate the number of reflecting elements in each sub-IRS $M_{\mathcal{C}}$, we rearrange (1) as

$$M_{\mathcal{C}} = \sqrt{\frac{M\lambda^2}{\pi \Delta_y \Delta_z}} = \lambda \sqrt{\frac{M_y M_z}{\pi \Delta_y \Delta_z}}. \tag{2}$$

We consider that the IRS is square-shaped, which means $M_y = M_z$, $\Delta_y = \Delta_z = \lambda/2^x$, $x \in \mathbb{N}$, and $\sqrt{1/\pi} \approx 1/2$. We rewrite (2) as follows:

$$M_{\mathcal{C}} = \frac{\lambda M_y}{\Delta_y} \sqrt{\frac{1}{\pi}} = \frac{2^x M_y}{2} = 2^{x-1} M_y. \tag{3}$$

Thus, the number of sub-IRSs can be obtained as

$$\mathcal{C} = \frac{M}{M_{\mathcal{C}}} = \frac{M_y}{2^{x-1}}. \tag{4}$$

There are multiple approaches to partition the IRS into $\mathcal{C}$ sub-IRSs, each with $M_{\mathcal{C}}$ reflecting elements. As shown in Fig. 2, the IRS is divided into $\mathcal{C}$ sub-IRSs, each with M_y columns and $M_{z,c} = 2^{x-1}$ rows. Moreover, our model is applicable for M_z rows and $M_y = 2^{x-1}$ columns.

C. 3D COORDINATE

In the 3D Cartesian coordinate system, the center locations of the cth sub-IRS $\mathrm{I}_{l,c}$, the AP, and the kth UE ($\mathrm{UE}_k \in \mathcal{K}$) are denoted as $\mathbf{p}_{l,c} = (x_{l,c}, y_{l,c}, z_{l,c})$, $\mathbf{p}_a = (x_a, y_a, z_a)$, and $\mathbf{p}_k = (x_k, y_k, z_k)$, respectively. Moreover, the distance from the AP to the center of $\mathrm{I}_{l,c}$ is written as

$$r_{a,l,c} \triangleq \sqrt{(x_a - x_{l,c})^2 + (y_a - y_{l,c})^2 + (z_a - z_{l,c})^2}. \tag{5}$$

Similarly, the distance from UE_k to $\mathrm{I}_{l,c}$ is written as

$$r_{l,c,k} \triangleq \sqrt{(x_k - x_{l,c})^2 + (y_k - y_{l,c})^2 + (z_k - z_{l,c})^2}. \tag{6}$$

For the sub-IRS $\mathrm{I}_{l,c}$, we denote $\check{\mathbf{p}}_{l,c,m} = (0, m_{y,c}\Delta_y, m_{z,c}\Delta_z)$ as the Cartesian coordinates of the mth reflecting element of the sub-IRS $\mathrm{I}_{l,c}$ with respect to its center, where $m_{y,c} = 0, \pm 1, \ldots, \frac{M_y - 1}{2}$, and $m_{z,c} = 0, \pm 1, \ldots, \frac{2^{x-1}-1}{2}$. Hence, the distance from the cth sub-array of the lth IRS's center to the mth reflecting element is calculated as

$$\check{r}_{l,c,m} = \sqrt{(m_{y,c}\Delta_y)^2 + (m_{z,c}\Delta_z)^2}. \tag{7}$$

We apply the following linear transformation to determine the Cartesian coordinates of the mth reflecting element (with respect to the origin).

$$\mathbf{p}_{l,c,m} = \mathbf{p}_{l,c} + \mathbf{R}_{l,c} \check{\mathbf{p}}_{l,c,m}, \tag{8}$$

where the rotation matrix $\mathbf{R}_{l,c} \in \mathbb{R}^{3\times 3}$ is applied to each IRS $\mathrm{I}_{l,c}$, $l \in \mathcal{L}$, $c \in \mathcal{C}$ to specify its own orientation. It is noted that $\mathbf{R}_{l,c}^{\mathsf{T}} = \mathbf{R}_{l,c}^{-1}$, $\mathbf{R}_{l,c}^{\mathsf{T}} \mathbf{R}_{l,c} = \mathbf{R}_{l,c} \mathbf{R}_{l,c}^{\mathsf{T}} = \mathbf{I}$, and $\det(\mathbf{R}_{l,c}) = 1$ [35]. Furthermore, the node's location in spherical coordinates is represented as (r, ψ, φ), where r is the radial distance, ψ is the elevation angle measured from the xy-plane upwards, and φ is the azimuth angle.

D. CHANNEL MODELING

In this section, we drive the cascaded channels through the sub-IRSs. We consider the downlink communication from the AP, equipped with N antennas, to a UE with a single antenna via the sub-IRSs. Furthermore, to evaluate the performance of the proposed optimization framework under ideal conditions, we assume perfect instantaneous channel state information (CSI). In practical deployments, due to model mismatch, channel estimation errors and outdated CSI may degrade achievable performance and affect convergence speed. Therefore, incorporating robust optimization techniques or statistical CSI-based formulations to account for channel uncertainty represents an important direction for future research. Let $\mathbf{g}_{n,l,c}^{\chi} \in \mathbb{C}^{M_c \times 1}$ denote the channel vector from the nth antenna of the AP (AP_n)-to-$\mathrm{I}_{l,c}$. Moreover, let $g_{n,(l,c,m)}^{\chi}$ represent the channel from AP_n-to-$\mathrm{I}_{l,c,m}$, which can be written as

$$g_{n,(l,c,m)}^{\chi} = \sqrt{\ell_{n,(l,c,m)}^{\chi}} e^{-j\omega_{\chi} r_{a,l,c}}, \tag{9}$$

where $\omega_\chi = \frac{2\pi}{\lambda_\chi}$ [m^{-1}] is the wavenumber, $\chi \in \{\mathbb{T}, \mathbb{M}\}$, $\omega_\mathbb{T}$ denotes the wavenumber at the THz operating frequency ($f_\mathbb{T}$), $\omega_\mathbb{M}$ denotes the wavenumber at the mmWave operating frequency ($f_\mathbb{M}$), and $\ell^\chi_{n,(l,c,m)}$ denotes the pathloss of the $\mathtt{AP}_n$-to-$\mathrm{I}_{l,c,m}$ link. The channel matrix from AP-to-$\mathrm{I}_{l,c}$ is represented as $\mathbf{G}^\chi_{l,c} = [\mathbf{g}^\chi_{1,l,c}, \ldots, \mathbf{g}^\chi_{n,l,c}, \ldots, \mathbf{g}^\chi_{N,l,c}] \in \mathbb{C}^{M_c \times N}$. The $\mathrm{I}_{l,c}$-to-$\mathtt{UE}_k$ channel at time slot t is denoted as $\mathbf{g}^\chi_{l,c,k}(t) \in \mathbb{C}^{M_c \times 1}$. Thus, $[\mathbf{g}^\chi_{l,c,k}]_{m_c}(t) = g^\chi_{(l,c,m),k}(t)$ is the $\mathrm{I}_{l,c,m}$-to-$\mathtt{UE}_k$ channel, which can be written as

$$g^\chi_{(l,c,m),k}(t) = \sqrt{\ell^\chi_{(l,c,m),k}(t)} e^{-j\omega_\chi r_{l,c,k}(t)}, \tag{10}$$

where $k \in \{i, j\}$, $\ell^\chi_{(l,c,m),k}(t)$ denotes the pathloss of the $\mathrm{I}_{l,c,m}$-to-$\mathtt{UE}_k$ link at time slot t. Let $\ell^\chi_k(t) = \ell^\chi_{n,(l,c,m)} \ell^\chi_{(l,c,m),k}(t)$ represent the overall pathloss of the considered cascaded downlink channel, which can be written as [18]

$$\ell^\chi_k(t) = \mathbb{G}_a \mathbb{G}_{l,c,m}(-\mathbf{r}_k) \mathbb{G}_{l,c,m}(\mathbf{r}_a) \mathbb{G}_k \times \frac{A^2 e^{-\kappa_{abs}(f_\chi)(r_{l,c,k}(t)+r_{a,l,c})}}{(4\pi r_{l,c,k}(t) r_{a,l,c})^2}, \tag{11}$$

where $\kappa_{abs}(f_\chi)$ denotes the molecular absorption coefficient, which models frequency-dependent f_χ atmospheric attenuation. The cascaded channel vector $\mathbf{h}^\chi_{l,c,k}(t) \in \mathbb{C}^{N \times 1}$ from AP-to-$\mathtt{UE}_k$ through $\mathrm{I}_{l,c}$ at time slot t can be expressed as [36]

$$\mathbf{h}^\chi_{l,c,k}(t) = (\mathbf{G}^\chi_{l,c})^{\mathsf{H}} \mathbf{\Theta}_{l,c} \mathbf{g}^\chi_{l,c,k}(t), \tag{12}$$

where $\mathbf{\Theta}_{l,c} \in \mathbb{C}^{M_y \times M_{z,c}}$ represents the PS matrix of $\mathrm{I}_{l,c}$, encompassing the amplitude and PS of each element of the (l, c)th sub-IRS. The PS matrix of the sub-IRS can be described as [37]

$$\mathbf{\Theta}_{l,c} = \mathrm{diag}([\kappa_{l,c,1} e^{j\theta_{l,c,1}}, \ldots, \kappa_{l,c,m} e^{j\theta_{l,c,m}}, \ldots, \kappa_{l,c,M_c} e^{j\theta_{l,c,M_c}}]), \tag{13}$$

where $\kappa_{l,c,m} \in [0, 1]$ and $\theta_{l,c,m} \in [0, 2\pi)$ represent the amplitude and the PS of $\mathrm{I}_{l,c,m}$, respectively. The channel matrix can be written as $\mathbf{H}^\chi_{l,c}(t) = [\mathbf{h}^\chi_{l,c,1}(t), .., \mathbf{h}^\chi_{l,c,k}(t), \ldots, \mathbf{h}^\chi_{l,c,K}(t)]^{\mathsf{H}} \in \mathbb{C}^{K \times N}$. Consider a precoding matrix $\mathbf{W} \in \mathbb{C}^{N \times K}$, where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k, \ldots, \mathbf{w}_K]$, and each $\mathbf{w}_k \in \mathbb{C}^{N \times 1}$ represents the precoding vector for $\mathtt{UE}_k$. This vector can be expressed as $\mathbf{w}_k = \sqrt{p_k}\widehat{\mathbf{w}}_k$, with p_k being the transmit power scaling factor and $\widehat{\mathbf{w}}_k$ the beamforming vector for $\mathtt{UE}_k$. Given power budget P_a, the power constraint is written as

$$\sum_{k=1}^{K} \|\mathbf{w}_k\|^2 = \sum_{i=1}^{I} \mathbf{w}_i + \sum_{j=1}^{J} \mathbb{I}_j \mathbf{w}_j \leq \mathrm{P}_a, \tag{14}$$

where $\mathbb{I}_j \in \{0, 1\}$, with 1 representing that $\mathtt{uUE}_j$ is transmitting and 0 showing that it is inactive.

E. SIGNAL MODELING AND DATA RATE ANALYSIS

Let $\mathbf{x} \in \mathbb{C}^N$ be the transmit signal vector from the AP to its UEs, where x_i and x_j are the messages intended for $\mathtt{eUE}_i \in \mathcal{I}$ and $\mathtt{uUE}_j \in \mathcal{J}$, respectively. Hence, the vector representing the transmitted signals from the AP can be expressed as [38]

$$\mathbf{x} = \sum_{i=1}^{I} \mathbf{w}_i x_i + \sum_{j=1}^{J} \mathbb{I}_j \mathbf{w}_j x_j = \sum_{i=1}^{I} \sqrt{p_i}\widehat{\mathbf{w}}_i x_i + \sum_{j=1}^{J} \mathbb{I}_j \sqrt{p_j}\widehat{\mathbf{w}}_j x_j. \tag{15}$$

The received signals at $\mathtt{eUE}_i \in \mathcal{I}$ through $\mathrm{I}_{l,c}$ at THz band can be expressed as

$$y^e_{l,c,i} = \underbrace{\sqrt{p_i}(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_i x_i}_{\text{desired signal}} + \underbrace{\sum_{i'=1, i' \neq i}^{I} \sum_{l=1}^{L} \sum_{c=1}^{C} \sqrt{p_{i'}}(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_{i'} x_{i'}}_{\text{interference}} + \underbrace{n_i}_{\text{noise}}, \tag{16}$$

where $n_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ_i^2 [W] at $\mathtt{eUE}_i \in \mathcal{I}$. The first part is the desired signal, which includes the intended signal for $\mathtt{eUE}_i$ reflected by the specific $\mathrm{I}_{l,c}$ allocated to it. The second part accounts for the interference, which arises from signals intended for all other users through all IRSs in the network, contributing to interference at $\mathtt{eUE}_i$. Finally, the third part is the noise present at $\mathtt{eUE}_i$. Furthermore, SINR $\gamma^e_{l,c,i}$ at $\mathtt{eUE}_i$ through $\mathrm{I}_{l,c}$, calculated using (16), can be expressed as

$$\gamma^e_{l,c,i} = \frac{p_i |(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_i|^2}{\sum_{i'=1, i' \neq i}^{I} \sum_{l=1}^{L} \sum_{c=1}^{C} p_{i'} |(\mathbf{h}^{\mathbb{T}}_{l,c,i})^{\mathsf{H}} \widehat{\mathbf{w}}_{i'}|^2 + \sigma_i^2}. \tag{17}$$

The received signals at $\mathtt{uUE}_j \in \mathcal{J}$ through $\mathrm{I}_{l,c}$ at mmWave band can be written as

$$y^u_{l,c,j} = \underbrace{\sqrt{p_j}(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_j x_j}_{\text{desired signal}} + \underbrace{\sum_{j'=1 j' \neq j}^{J} \sum_{l=1}^{L} \sum_{c=1}^{C} \mathbb{I}_{j'} \sqrt{p_{j'}}(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_{j'} x_{j'}}_{\text{interference}} + \underbrace{n_j}_{\text{noise}}, \tag{18}$$

where the first part is the desired signal, which includes the intended signal for $\mathtt{uUE}_j$ reflected by the specific $\mathrm{I}_{l,c}$ allocated to it. The second part accounts for the interference, which arises from signals intended for all other users through all IRSs in the network, contributing to interference at $\mathtt{uUE}_j$. Finally, the third part $n_j \sim \mathcal{CN}(0, \sigma_j^2)$ is the AWGN with zero mean and variance σ_j^2 [W] at $\mathtt{uUE}_j \in \mathcal{J}$. SINR $\gamma^u_{l,c,j}$ can be formulated as

$$\gamma^u_{l,c,j} = \frac{p_j |(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_j|^2}{\sum_{j'=1, j' \neq j}^{J} \sum_{l=1}^{L} \sum_{c=1}^{C} p_{j'} |(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_{j'}|^2 + \sigma_j^2}. \tag{19}$$

1) DATA RATE OF FeMBB

The achievable rate of $\texttt{eUE}_i \in \mathcal{I}$ through $\mathrm{I}_{l,c}$ can be expressed using Shannon's data rate formulation as

$$R^e_{l,c,i} = \log_2(1 + \frac{p_i |(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_i|^2}{\sum_{i' \neq i}^{I} \sum_{l=1}^{L} \sum_{c=1}^{C} p_{i'} |(\mathbf{h}^{\mathbb{T}}_{l,c,i})^{\mathsf{H}} \widehat{\mathbf{w}}_{i'}|^2 + \sigma_i^2}). \tag{20}$$

Furthermore, the data rate of $\texttt{eUE}_i \in \mathcal{I}$ can be written as

$$R^e_i = \sum_{l=1}^{L} \sum_{c=1}^{C} R^e_{l,c,i} [\Psi]_{(l,c),i}, \tag{21}$$

where $[\Psi]_{(l,c),i} \in \{0, 1\}$ is the sub-IRS allocation indicator, which is set to 1 when $\mathrm{I}_{(l,c)}$ is allocated to $\texttt{eUE}_i$, and set to 0 otherwise.

2) DATA RATE OF eURLLC

In eURLLC, the well-known Shannon's data rate formula for long block-length wireless communications is inaccurate. Thus, we use the finite block length formula to calculate the achievable rate of eURLLC UEs, which can be expressed as [39] and [40]

$$\begin{aligned} R^u_{l,c,j} &= \log_2\left(1 + \gamma^u_{l,c,j}\right) - \sqrt{\frac{V_j}{\zeta_j}} Q^{-1}(\epsilon_j) + \frac{O(\log \zeta_j)}{\zeta_j} \\ &\approx \log_2\left(1 + \gamma^u_{l,c,j}\right) - \sqrt{\frac{V_j}{\zeta_j}} Q^{-1}(\epsilon_j), \end{aligned} \tag{22}$$

where ζ_j, ϵ_j, and V_j denote the packet length, error probability, and channel dispersion, respectively. The channel dispersion is defined as $V_j = 1 - (1 + \gamma^u_{l,c,j})^{-2}$. Furthermore, $Q^{-1}(\epsilon_j)$ is the inverse Q-function[3] of ϵ_j. To obtain the error probability at $\texttt{uUE}_j$ which is paired with a specific $\texttt{eUE}_i$ and punctured the allocated sub-IRS $\mathrm{I}_{l,c}$, one can reformulate (22) by simple manipulation, which yields the following:

$$\begin{aligned} R^u_{l,c,j}[\Upsilon]_{i,j} &= \log_2(1 + \gamma^u_{l,c,j}[\Upsilon]_{i,j}) - \sqrt{\frac{V_j}{\zeta_j}} Q^{-1}(\epsilon_j[\Upsilon]_{i,j}), \\ Q^{-1}(\epsilon_j[\Upsilon]_{i,j}) &= \sqrt{\frac{\zeta_j}{V_j}} \left(\log_2(1 + \gamma^u_{l,c,j}[\Upsilon]_{i,j}) - R^u_{l,c,j}[\Upsilon]_{i,j}\right), \\ \epsilon_j[\Upsilon]_{i,j} &= Q\left(\sqrt{\frac{\zeta_j}{V_j}} (\log_2(1 + \gamma^u_{l,c,j}[\Upsilon]_{i,j}) - R^u_{l,c,j}[\Upsilon]_{i,j})\right), \end{aligned} \tag{23}$$

where $[\Upsilon]_{i,j} \in \{0, 1\}$ is the allocation indicator, which is set to 1 when $\texttt{eUE}_i$ is allocated to $\texttt{uUE}_j$, and to 0 otherwise.

F. IRS REFLECTION MATRIX CONFIGURATIONS

The reflection matrix configuration is an important characteristic of the IRS. The reflection matrix combines the amplitude and PS of each element of the sub-IRS. In order to maximize the power gain, the PS induced by $\mathrm{I}_{l,c,m}$ is considered as an ideal PS configuration, and the amplitude of each sub-IRS element is set to 1, as follows [41]

$$\phi^\star_{l,c,m} = \frac{2\pi}{\lambda}(r_{a,l,c} + r_{l,c,k})), \quad \forall l, c, m. \tag{24}$$

$$\kappa^\star_{l,c,m} = 1, \quad \forall l, c, m. \tag{25}$$

Therefore, the optimal reflection matrix configuration at $\mathrm{I}_{l,c}$ is determined as

$$\begin{aligned} \mathbf{\Theta}^\star_{l,c} = \mathrm{diag}[&\kappa^\star_{l,c,1} e^{j\phi^\star_{l,c,1}}, \ldots, \kappa^\star_{l,c,m} e^{j\phi^\star_{l,c,m}}, \\ &\ldots, \kappa^\star_{l,c,M_c} e^{j\phi^\star_{l,c,M_c}}]. \end{aligned} \tag{26}$$

G. NETWORK MOBILITY MODEL

Mobility models aim to capture the displacement of UEs as they traverse different locations. These models are categorized into two main types: individual movement models and group movement models [42]. For the purpose of our analysis, we use the random waypoint mobility (RWM) model for both $\texttt{eUE}$ and $\texttt{uUE}$, where UEs move randomly within a designated area. This model is suitable for simulating IoE networks with freely moving UEs. In the RWM model, the velocity and direction of $\texttt{UE}_k$ at time t are represented as $v_k(t) \in [v_{min}, v_{max}]$ and $\phi_k(t) \in [0, 2\pi]$, respectively, and can be determined by a memoryless random process. Furthermore, for a given interval of data collection T, we have

$$\begin{aligned} \sin(\phi_k(t)) &= \frac{y_k(t+T) - y_k(t)}{T v_k(t)}, \\ \cos(\phi_k(t)) &= \frac{x_k(t+T) - x_k(t)}{T v_k(t)}. \end{aligned} \tag{27}$$

Let $\alpha_a \in (-1, 1)$ represent the angle deviation factor and $\alpha_s \in (-1, 1)$ denote the speed deviation factor. The UEs' mobility between times t_1 and t_4 is illustrated in Fig. 3.

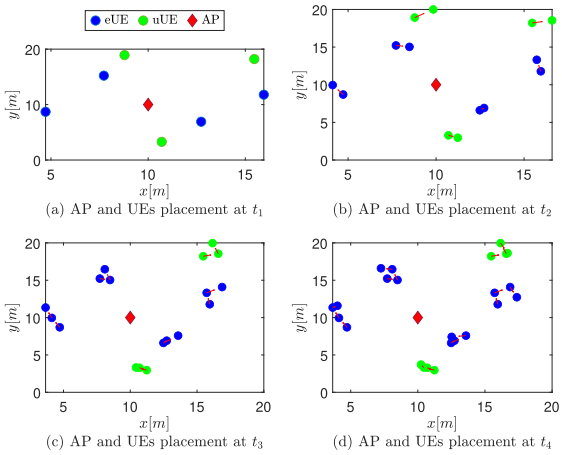


FIGURE 3. UEs and the AP placement at t_1 to t_4 during UEs' movement.

III. OPTIMIZATION PROBLEM FORMULATION

In this section, we formulate the MOP to maximize the sum rate of the FeMBB UEs and the reliability of the eURLLC UEs. The optimization variables are the FeMBB UEs' allocation matrix Ψ, the eURLLC UEs' allocation matrix Υ, the

[3] As usual, $Q(x) = \frac{1}{2\pi} \int_x^\infty e^{-v/2}\, dv$

AP's transmit powers to the FeMBB and eURLLC UEs, $\mathbf{P^e}$ and $\mathbf{P^u}$, respectively. The FeMBB UEs' sum rate and the eURLLC UEs' reliability can be jointly maximized via the following MOP:

$$2\mathbf{P}: \max_{\mathbf{P}^e,\Psi,\mathbf{P}^u,\Upsilon} \Big(\sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} R^e_{l,c,i}[\Psi]_{(l,c),i}\Big), \Big(-\sum_{j=1}^{J}\epsilon_j[\Upsilon]_{i,j}\Big) \tag{28a}$$

$$\text{s.t.} \quad R^e_{l,c,i} \geq R_{\min}, \quad \forall i \in I, \tag{28b}$$
$$p_i \geq 0, \quad \forall i \in I, \tag{28c}$$
$$\textstyle\sum_{i\in I} p_i \leq P^e_{\max}, \tag{28d}$$
$$\Psi_{l,c,i} \in \{0,1\}, \quad \forall i \in I, \tag{28e}$$
$$\textstyle\sum_{i\in I} \Psi_{l,c,i} \leq C, \quad \forall c \in C, \tag{28f}$$
$$\textstyle\sum_{l\in L} \Psi_{l,c,i} \leq \varsigma_i, \quad \forall i \in I, \tag{28g}$$
$$p_j \geq 0, \quad \forall j \in J, \tag{28h}$$
$$\textstyle\sum_{j\in J} p_j \leq P^u_{\max}, \tag{28i}$$
$$\Upsilon_{i,j} \in \{0,1\}, \quad \forall j \in J, \tag{28j}$$
$$\textstyle\sum_{j\in J} \Upsilon_{i,j} \leq 1, \quad \forall i \in I, \tag{28k}$$
$$\textstyle\sum_{i\in I} \Upsilon_{i,j} \leq 1, \quad \forall j \in J, \tag{28l}$$
$$\epsilon_j \leq \epsilon_{\max}, \quad \forall j \in \mathcal{J}, \tag{28m}$$
$$\tau_j \leq \tau_{\max}, \quad \forall j \in J, \tag{28n}$$

where $R_{\min}$ is the minimum required data rate for the FeMBB UEs, ς_i is the quote of $\in$UE$_i$, $\epsilon_{\max}$ denotes the maximum allowed error probability for the eURLLC UEs, and $P^e_{\max}$ and $P^u_{\max}$ represent the maximum power budget. Constraint (28b) imposes a minimum data rate for each FeMBB UE while meeting the eURLLC UEs' reliability. Furthermore, Constraints (28c) and (28d) ensure that the transmit powers are always positive and that, for the FeMBB UEs, the total power does not exceed the AP's power budget, respectively. Moreover, Constraints (28e), (28f), and (28g) ensure many-to-many matching within the quota. The power constraints on eURLLC UEs are provided in (28h) and (28i). Constraints (28j), (28k), and (28l) guarantee the one-to-one matching of eURLLC UEs. Constraint (28m) restricts the error probability to meet the eURLLC UEs' reliability, while Constraint (28n) guarantees that the eURLLC UEs are served within the latency requirement. The delay requirement of uUE$_j$ is ensured as $\tau_j \leq C_j/R_{l,c,j}$, where C_j represents the total transmitted bits per packet for uUE$_j$. Furthermore, as usual for MOP, the solution is a set of local solutions, rather than a global optimal solution. To solve the complex MOP problem, current solutions usually transform it into an SOP to be more tractable and computationally efficient. The scalarization and weighted sum method is proposed to convert the MOP to a SOP; thus, the MOP in (28) can be written as follows:

$$\mathbf{P1}: \max_{\mathbf{P}^e,\Psi,\mathbf{P}^u,\Upsilon} \delta \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} R^e_{l,c,i}[\Psi]_{(l,c),i} + (1-\delta)\Big(-\sum_{j=1}^{J}\epsilon_j[\Upsilon]_{i,j}\Big)$$
$$\text{s.t.} \quad (28b) \text{ - } (28n), \tag{29}$$

where $\delta \in \{0, 1\}$ denotes the weight parameter. Of note, the FeMBB sum rate and the eURLLC error probability represent heterogeneous performance metrics with different physical meanings and numerical scales. To ensure a balanced aggregation in the weighted-sum formulation, the weighting parameter $\delta \in [0, 1]$ is selected to appropriately scale their relative contributions in the objective function. In particular, the eURLLC error probability is inherently bounded due to the reliability constraint, while the FeMBB sum rate remains within a predictable range under the considered system configuration. Therefore, the weighting parameter effectively compensates for scale differences between the two metrics. Moreover, constant normalization factors, if explicitly introduced, can be equivalently absorbed into the weighting coefficients without altering the structure of the optimization problem. This formulation ensures a meaningful trade-off between throughput and reliability despite their different units and numerical ranges.

IV. PROPOSED RESOURCE ALLOCATION SOLUTION

Problem **P1** is a MINLP problem, which is generally very hard to solve and is an NP-hard problem. Therefore, we decompose **P1** into sub-problems to find a sub-optimal solution. Specifically, we divide **P1** into two sub-problems for FeMBB UEs RA and eURLLC UEs RA.

A. FeMBB UEs RESOURCE ALLOCATION

For any given eURLLC UEs resources Υ and $\mathbf{P^u}$, we set $\delta = 1$, thus, the optimization problem **P1** in (29) can be expressed as

$$\mathbf{P2}: \max_{\mathbf{P}^e,\Psi} \textstyle\sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} R^e_{l,c,i}[\Psi]_{(l,c),i}$$
$$\text{s.t.} \quad (28b) \text{ - } (28g). \tag{30}$$

The FeMBB RA optimization problem in **P2** is still an NP-hard problem. Therefore, we decompose **P2** into two sub-problems, FeMBB UEs-sub-IRSs association and FeMBB UEs power allocation.

1) FeMBB UEs-SUB-IRSs ASSOCIATION

To maximize the sum rate of FeMBB UEs, we optimize the association matrix for a given power allocation matrix. The corresponding optimization problem is defined as:

$$\mathbf{P2.1}: \max_{\Psi} \textstyle\sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} R^e_{l,c,i}[\Psi]_{(l,c),i}$$
$$\text{s.t.} \quad (28b), (28e), (28g). \tag{31}$$

Now, sub-problem (31) turns into a resource scheduling problem, which we can solve using the proposed matching scheme. In our system model, we considered $L \times C > K$ number of sub-IRSs. Therefore, we propose a many-to-many-matching scheme; in this matching, we allocate multiple IRSs

Algorithm 1 Proposed Many-to-Many Matching-Based FeMBB UE Association Algorithm for **P2.1**

Input: $l \in L$, $c \in C$, $i \in I$, $R^{e}_{l,c,i}$, $[\Upsilon]^{e}_{i,l,c}$, $[\Upsilon]^{\mathrm{I}}_{l,c,i}$, set of unmatched FeMBB UEs Π, FeMBB UE quota ς, $[\Psi]_{(l,c),i}$

Initialize $[\Upsilon]^{e}_{i,l,c} = \emptyset$, $[\Upsilon]^{\mathrm{I}}_{l,c,i} = \emptyset$, $[\Psi]_{(l,c),i} = \emptyset$;

Step 1: Preference Matrix Configuration:

for $i \in I$ **do**

  $R^{e}_{l,c,i}$, $\forall i \in I$ and store in $[\mathbf{R}]^{e}_{l,c,i}$

$[[\mathbf{R}]^{e}_{l,c,i}, [\Upsilon]^{e}_{i,l,c}]$ = sort $([\mathbf{R}]^{e}_{l,c,i}$,2,descend)

$[[\mathbf{R}]^{\mathrm{I}}_{l,c,i}, [\Upsilon]^{\mathrm{I}}_{l,c,i}]$ = sort $([[\mathbf{R}]^{e}_{l,c,i}]^{\mathsf{T}}$,2,descend)

Step 2: FeMBB UE Association:

while *($\Pi \neq \emptyset \cup$ FeMBB UEs not rejected by all IRSs)* **do**

  if $(\varsigma_i > \varsigma)$

    No more IRS allocates to $\mathtt{eUE}_i$

  else

    for $\mathtt{eUE}_{i'} \in \Pi$ **do**

      Target the highest priority sub-IRS in $[\Upsilon]^{e}_{i,l,c}$

      Element of $[\Psi]_{(l,c),i} = 1$

    for $\mathrm{I}_{l,c} \in l, c$ **do**

      if $(\mathrm{I}_{l,c} \notin [\Psi]_{(l,c),i})$

        $[\Psi]_{(l,c),i} \leftarrow [\Psi]_{(l,c),i} \cup (\mathtt{eUE}_{i'}, \mathrm{I}_{l,c})$,

        $\varsigma_{i'} = \varsigma_{i'} + 1$

      else if $(R^{\mathrm{I}}_{l,c,i'} > R^{\mathrm{I}}_{l,c,i})$

        Break the existing allocation,

        $[\Psi]_{(l,c),i} \leftarrow [\Psi]_{(l,c),i} \cup (\mathtt{eUE}_{i'}, \mathrm{I}_{l,c})$

        $\varsigma_{i'} = \varsigma_{i'} + 1$

      else Reject $\mathtt{eUE}_{i'}$ and continue with the existing allocation

Output: FeMBB UE Association matrix Ψ^{*}

to assist the FeMBB UEs. Each $\mathtt{eUE}_i \in I$ calculates its achievable rate $R^{e}_{l,c,i}$ over the set of sub-IRSs using (20). Furthermore, based on the calculated data rates, each FeMBB UE defines its priority relation and prepares the priority matrix $\Upsilon^{e}_{i,l,c}$. The priority matrix of the sub-IRSs can be represented as $\Upsilon^{\mathrm{I}}_{l,c,i}$, where the priority matrix configuration process is shown in the first step of Algorithm 1. In the original one-to-one matching problem, one sub-IRS is allocated to only one UE. However, in our proposed matching, multiple sub-IRSs can be allocated to one FeMBB UE. We define the quote ς_i as the quote of $\mathtt{eUE}_i$, which is the maximum number of sub-IRSs that can be allocated to $\mathtt{eUE}_i$. The sub-IRS allocation process is shown in the second step of Algorithm 1.

2) FeMBB UEs POWER ALLOCATION

At the beginning of the slot, all the power is allocated to the FeMBB UEs. Thus, the power-allocation sub-problem for a given $\Psi^{\star}$ can be formulated as

$$\mathbf{P2.2}: \max_{\mathbf{P}^{e}} \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} R^{e}_{l,c,i}[\Psi]_{(l,c),i} \quad \text{s.t.} \quad (28b)\text{ - }(28d). \tag{32}$$

Algorithm 2 FeMBB UEs Power Allocation Algorithm for **P2.2**

Input: I, FeMBB UE power allocation matrix $\mathbf{P}^{e} = [p_i]$, convergence threshold $\xi^{e} > 0$

Initialize: $p_i(0) = \mathbf{P}^{e}_{\max}/I$, $t = 0$

for $i \in I$ **do**

  while $(p_i(t) - p_i(t-1) > \xi^{e})$ **do**

    $\hat{p}_i \leftarrow p_i(t)$ and $t = t + 1$

    while *($\hat{p}_i$ is not converged)* **do**

      Obtain μ^{e}, ϱ^{e}_i, and υ^{e}_i via bisection on $P^{e}_{\max}$ in (34)

      $p_i = \left[\frac{1+\upsilon^{e}_i}{\mu^{e}+\varrho^{e}_i} - \frac{\iota_i(t)+\sigma^2_i}{\sum_{l\in L}\sum_{c\in C}[\Psi]_{(l,c),i}|(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}\right]^{+}$

      Normalize to meet the power budget constraints

      $p_i(t) \leftarrow \hat{p}_i$

Output: FeMBB UE Power allocation matrix $\mathbf{P}^{e\star}$

Theorem 1: The closed-form expressions for the optimal power allocation of $\mathtt{eUE}_i$ and power budget are given as, respectively

$$p_i = \left[\frac{1+\upsilon^{e}_i}{\mu^{e}+\varrho^{e}_i} - \frac{\iota_i(t)+\sigma^2_i}{\sum_{l\in L}\sum_{c\in C}[\Psi]_{(l,c),i}|(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}\right]^{+}, \tag{33}$$

$$P^{e}_{\max} = \sum_{i=1}^{I}\left[\frac{1+\upsilon^{e}_i}{\mu^{e}+\varrho^{e}_i} - \frac{\iota_i(t)+\sigma^2_i}{\sum_{l\in L}\sum_{c\in C}[\Psi]_{(l,c),i}|(\mathbf{h}^{\mathbb{T}}_{l,c,i}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}\right]^{+}, \tag{34}$$

where $[x]^{+} = \max\{x, 0\}$ and μ^{e}, ϱ^{e}_i, and υ^{e}_i are the Lagrangian multipliers. The value of μ^{e} is selected to ensure that $\sum_{i\in I} p^{e}_i \leq P_{\max}$, ϱ^{e}_i is selected to avoid a non-negative value for p^{e}_i, and υ is selected to ensure the minimum required data rate is satisfied.

Proof: The proof is provided in Appendix A. ∎

We use Theorem 1 to determine the optimal p_i and then apply the water-filling (WF) method iteratively across all UEs. The iterative WF method for (32) is presented in Algorithm 2. The process of solving Lagrange multipliers μ^{e}, ϱ^{e}_i, and υ^{e}_i in Algorithm 2 employs a bisection search technique to ensure that the total power allocation satisfies constraint P^{e}_{max}. This method systematically adjusts μ^{e} by iterating within predefined lower and upper bounds until the allocated power meets the maximum allowable power with sufficient precision. Multipliers ϱ^{e}_i and υ^{e}_i are updated iteratively to satisfy individual constraints, such as the data rate requirements and per-UE power constraints, ensuring alignment with μ^{e}.

B. eURLLC UEs RESOURCE ALLOCATION

For any given FeMBB resources Ψ and $\mathbf{P^e}$, we set $\delta = 0$; thus, the optimization problem **P1** in (29) can be expressed as

$$\mathbf{P3}: \max_{\mathbf{P}^u, \Upsilon} \; -\sum_{j=1}^{J} \epsilon_j [\Upsilon]_{i,j} \quad \text{s.t. (28}h\text{)-(28}n\text{).} \tag{35}$$

The eURLLC RA optimization problem in **P3** is also an NP-hard problem. Thus, we decompose **P3** into two sub-problems, eURLLC UEs association and eURLLC power UEs allocation.

1) eURLLC UEs-FeMBB UEs ASSOCIATION

To maximize reliability of eURLLC UEs, we optimize the association matrix for the given power allocation. The corresponding optimization problem is defined as follows:

$$\mathbf{P3.1}: \max_{\Upsilon} \; -\sum_{j=1}^{I} \epsilon_j [\Upsilon]_{i,j} \quad \text{s.t. (28}j\text{)-(28}n\text{).} \tag{36}$$

Sub-problem **P3.1** in (36) turns into a resource pairing problem, which we can solve using the proposed one-to-one matching-based algorithm. Each FeMBB UE, $\texttt{eUE}_i$ computes its achievable rate, establishes its priority relation, and constructs priority matrix $\Gamma^e_{i,j}$. Moreover, each eURLLC UE, $\texttt{uUE}_j$ computes its error rate, establishes its priority relation, and constructs priority matrix $\Gamma^u_{j,i}$, as shown in step 1 of Algorithm 3. Each eURLLC UE is potentially allocated to a single FeMBB UE in the one-to-one matching process. Algorithm 3 is depicted as a sequence of attempts by eURLLC UEs to form connections with FeMBB UEs. During this process, each FeMBB UE may either be matched with an eURLLC UE or remain unpaired. eURLLC UEs initiate the association process by proposing to the highest-priority FeMBB UE in their priority list. This proposal continues until an eURLLC UE is either accepted or rejected by all FeMBB UEs. Once rejected by a FeMBB UE, an eURLLC UE can no longer propose to the same FeMBB UE. If an eURLLC UE proposes to an unpaired FeMBB UE, that FeMBB UE immediately accepts the proposal. However, if the eURLLC UE proposes to a FeMBB UE that is already engaged, that FeMBB UE compares the new and current eURLLC UE and selects the one with the highest priority. If the FeMBB UE favors the current eURLLC UE, the new proposal is rejected. However, if the FeMBB UE prefers the new proposal, it breaks its engagement with the current eURLLC UE and accepts the new one. This process iterates until all eURLLC UEs are paired or all possibilities are exhausted, as outlined in step 2 of Algorithm 3.

Remark 1: The proposed many-to-many matching-based algorithm and one-to-one matching-based algorithm terminate in polynomial time. Specifically, if there are I FeMBB UEs, J eURLLC UEs, and L IRSs with C sub-IRSs, then Algorithm 1 and Algorithm 3 terminate after a maximum of $I \times L \times C$ and $I \times J$ iterations, respectively.

Algorithm 3 Proposed One-to-One Matching-Based eURLLC UE Association Algorithm for **P3.1**

Input: I, J, $[\Gamma]^e_{i,j}$, $[\Gamma]^u_{j,i}$, set of unmatched eURLLC UEs Π^u, $[\Upsilon]_{i,j}$

Initialize $[\Gamma]^e_{i,j} = \emptyset$, $[\Gamma]^u_{j,i} = \emptyset$, $[\Upsilon]_{i,j} = \emptyset$;

Step 1: Preference Matrix Configuration:

$[[\mathbf{R}]^e_{l,c,i}, [\Gamma]^e_{i,j}]$ = sort $([\mathbf{R}]^e_{l,c,i}, 2, \text{descend})$

$[[\epsilon]^u_j, [\Gamma]^u_{j,i}]$ = sort $([\epsilon]^u_j, 2, \text{ascend})$

Step 2: eURLLC UE Association:

while *($\Pi^u \neq \emptyset \cup$ eURLLC UEs not rejected by all FeMBB UEs)* **do**

  for $\texttt{uUE}_{j'} \in \Pi^u$ **do**

    Target the highest priority $\texttt{eUE}$ in $[\Gamma]^u_{j,i}$

    Element of $[\Upsilon]_{i,j} = 1$

  for $\texttt{eUE}_i \in I$ **do**

    if $(\texttt{eUE}_i \notin [\Upsilon]_{i,j})$

    $[\Upsilon]_{i,j} \leftarrow [\Upsilon]_{i,j} \cup (\texttt{uUE}_{j'}, \texttt{eUE}_i)$

    else if Priority of new proposal is higher than the existing pair

    Break the existing pairing

    $[\Upsilon]_{i,j} \leftarrow [\Upsilon]_{i,j} \cup (\texttt{uUE}_{j'}, \texttt{eUE}_i)$

    else Reject $\texttt{uUE}_{j'}$ and continue with the existing pairing

Output: eURLLC UE Association matrix Υ^*

Proof: In each iteration of the algorithm, every unmatched FeMBB UE puts forward a proposal to the previously unexplored sub-IRS. Subsequently, the sub-IRS either accepts or declines the proposal based on its status and preferences, as outlined in Step 2 of Algorithm 1. Considering I FeMBB UEs and $L \times C$ total sub-IRSs, there can be a maximum of $I \times L \times C$ possible proposals within the proposed algorithm. Similarly, we can prove that Algorithm 3 terminates after a maximum of $I \times J$ iterations. ∎

2) eURLLC UEs POWER ALLOCATION

To achieve maximum reliability for eURLLC UEs, we aim to optimize the transmit power for a given eURLLC UE association matrix Υ^*. This leads to the following optimization problem:

$$\mathbf{P3.2}: \max_{\mathbf{P}^u} \; -\sum_{j=1}^{J} \epsilon_j [\Upsilon]_{i,j} \quad \text{s.t.} \quad (28h), (28i). \tag{37}$$

To maximize reliability of eURLLC UEs for the given allocation matrix, we need to minimize the error probability. Using (37) and (23), yields

$$\max_{\mathbf{P}^u} -\sum_{j=1}^{J} \epsilon_j = \min_{\mathbf{P}^u} \sum_{j=1}^{J} \epsilon_j \approx \max_{\mathbf{P}^u} \sum_{j=1}^{J} \sqrt{\frac{\zeta_j}{V_j}} \left(\log_2(1 + \gamma^u_{l,j}) - R^u_{l,c,j} \right). \tag{38}$$

Algorithm 4 eURLLC UEs Power Allocation Algorithm for **P3.2**

Input: J, eURLLC UE power allocation matrix $\mathbf{P}^u = [p_j]$, convergence threshold $\xi^u > 0$
Initialize: $p_j(0) = \mathbf{P}^u_{\max}/J$, $n = 0$
for $j \in J$ **do**
  while $(p_j(t) - p_j(t-1) > \xi^u)$ **do**
    $\hat{p}_j \leftarrow p_j(t)$
    $t = t + 1$
    while *($\hat{p}_j$ is not converged)* **do**
      Obtain μ^u and ϱ^u_j via bisection on $P^u_{\max}$ in (40)
      $p_j = \left[\sqrt{\frac{\zeta_j}{V_j}}\left(\frac{1}{\mu^u+\varrho^u_i}\right) - \frac{\iota_j(t)+\sigma_j^2}{|(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}\right]^+$.
      Normalize to meet the power budget constraints
      $p_j(t) \leftarrow \hat{p}_j$
Output: eURLLC UE Power allocation matrix $\mathbf{P}^{u\star}$

Theorem 2: The closed-form expressions for the optimal power allocation of $\mathtt{uUE}_j$ and power budget are given as

$$p_j = \left[\sqrt{\frac{\zeta_j}{V_j}}\left(\frac{1}{\mu^u + \varrho^u_i}\right) - \frac{\iota_j(t) + \sigma_j^2}{|(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}\right]^+, \tag{39}$$

$$P^u_{\max} = \sum_{j=1}^{J}\left[\sqrt{\frac{\zeta_j}{V_j}}\left(\frac{1}{\mu^u + \varrho^u_i}\right) - \frac{\iota_j(t) + \sigma_j^2}{|(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}\right]^+, \tag{40}$$

respectively. Here, μ^u and ϱ^u_i are the Lagrangian multipliers for the total power constraint and the non-negativity constraints on p_j.

Proof: The proof is provided in Appendix B. ∎

The optimal p^u_j is obtained using Theorem 2 and then the WF method is iterated over all UEs. The WF method proposed for (37) is detailed in Algorithm 4. The process of solving Lagrange multipliers μ^u and ϱ^u_j in Algorithm 4 employs a bisection search technique to ensure that the total power allocation satisfies the constraint P^u_{max}. This method systematically adjusts μ^u by iterating within predefined lower and upper bounds until the allocated power meets the maximum allowable power with sufficient precision. Multiplier ϱ^u_j is iteratively updated to satisfy individual constraints, such as the reliability requirements and per-UE power constraints, ensuring alignment with μ^u.

C. COMPUTATIONAL COMPLEXITY ASSESSMENT

To achieve the optimal FeMBB-sub-IRS pairing, the ES method involves evaluating all possible combinations and selecting the one that maximizes the sum rate. The computational complexity of this approach is $\mathcal{O}((L \times C)!)$. In contrast, the computational complexity of the proposed matching-based method is $\mathcal{O}(I \times L \times C)$. To assess the computational complexities of these schemes, we analyze their logarithmic complexities. The logarithmic complexity of the ES method is $\mathcal{O}(\ln((L \times C)!))$. By applying logarithmic properties and Stirling's formula ($\ln(n!) = n\ln(n)$), this can be simplified to $\mathcal{O}((L \times C)\ln(L \times C))$. On the other hand, the logarithmic complexity of the proposed scheme is $\mathcal{O}(\ln(I) + \ln(L) + \ln(C))$, which is significantly lower than that of the ES method. Similarly, we can calculate the computational complexity of the FeMBB-URLLC matching algorithm. For a detailed comparison of matching complexities of the proposed scheme and the ES scheme, see Table 2.

TABLE 2. Complexity comparison of different schemes.

K	L	C	No. of iterations for ES	No. of iterations for proposed scheme
4	4	4	24	16
8	8	4	40,320	64
10	10	4	3,628,800	100

V. SIMULATION RESULTS AND DISCUSSION

In this section, we present simulation results to evaluate the performance of our proposed algorithm and compare it to existing schemes. Our study employs MATLAB to implement the algorithms, with each data point derived from the average of 10^6 independent trials. In the simulated scenario, the carrier frequency and channel bandwidth for the THz band are $f_{\mathbb{T}} = 300$ GHz and 10 GHz, respectively [43]. Similarly, the carrier frequency and channel bandwidth for the mmWave band are $f_{\mathbb{M}} = 28$ GHz and 1 GHz, respectively [44]. Additionally, the dimension of each element in the IRS is $\lambda/2$ meters, where λ represents the wavelength. The simulation parameters, outlined in Table 3, are configured based on prior research works [18], [19], [43], [44]. We compare the proposed scheme with the following methods:

- **Exhaustive Search (ES):** The ES method uses the priority matrix to evaluate all possible pairings and select the combination that maximizes the objective function.
- **Greedy Search (GS):** In the GS method [45], [46], each UE selects the IRS with the highest priority. When multiple proposals are made to the same IRS, it is randomly assigned to one of the UEs.
- **Random Search (RS):** The RS scheme [47] involves randomly selecting the UE-IRS association matrix.

A. PERFORMANCE EVALUATION OF FeMBB

Fig. 4 shows the FeMBB sum rate for different schemes relative to the transmit power of the AP. It is observed that the sum rate increases steadily with higher transmission power for all schemes. Remarkably, the sum rate achieved by the proposed algorithm is nearly identical to that of the ES scheme. Additionally, the proposed algorithm outperforms the GS and RS methods. Fig. 5 shows the effect of the number of sub-IRSs on the FeMBB sum rate for various schemes under a fixed eURLLC load. The FeMBB sum rate achieved with the proposed algorithm is close to the one obtained with the ES scheme. Furthermore, the results highlight the

TABLE 3. Simulation parameters.

Parameters	Values
Number of AP antennas	64
Number of UEs	5-20
Number of IRS elements	100×100 [19]
THz frequency ($f_{\mathbb{T}}$) [GHz]	300 [19]
mmWave frequency ($f_{\mathbb{M}}$) [GHz]	28 [44]
Absorption coefficient at $f_{\mathbb{T}}$ $[m^{-1}]$	0.0033 [18]
Absorption coefficient at $f_{\mathbb{M}}$ $[m^{-1}]$	0.000014 [48]
Transmit power [dBm]	25 [18]
Noise power density for THz band [dBm/Hz]	-174 [43]
Noise power density for mmWave band [dBm/Hz]	-134 [44]
Channel BW in THz band [GHz]	10 [43]
Channel BW in mmWave band [GHz]	1
Noise figure [dB]	10 [19]
Side length of IRS elements $[m]$	$\lambda/2$

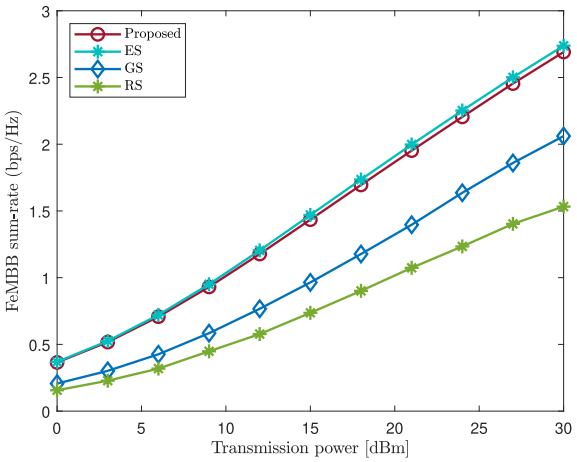


FIGURE 4. FeMBB sum-rate vs. transmit power.

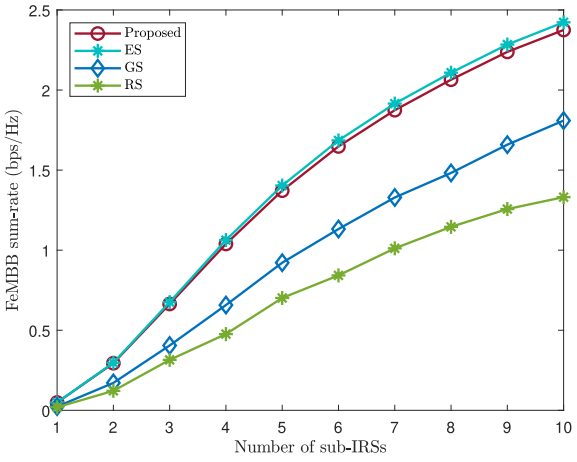


FIGURE 5. FeMBB sum-rate vs. number of sub-IRSs.

proposed scheme's superior performance as compared to the GS and RS schemes. This enhancement is attributed to the effective association of the sub-IRSs, which improves the channel gain. Specifically, with a configuration of 10

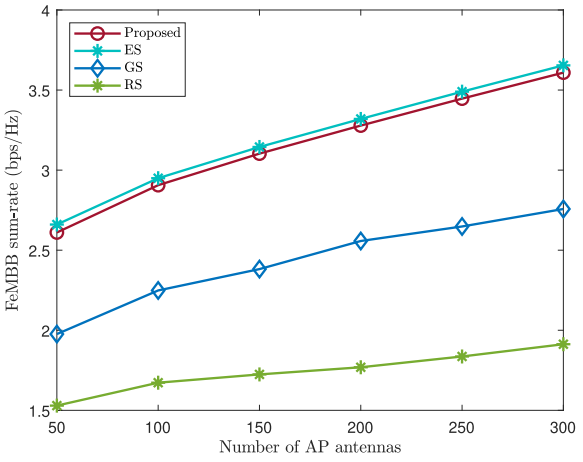


FIGURE 6. FeMBB sum-rate vs. number of AP antennas.

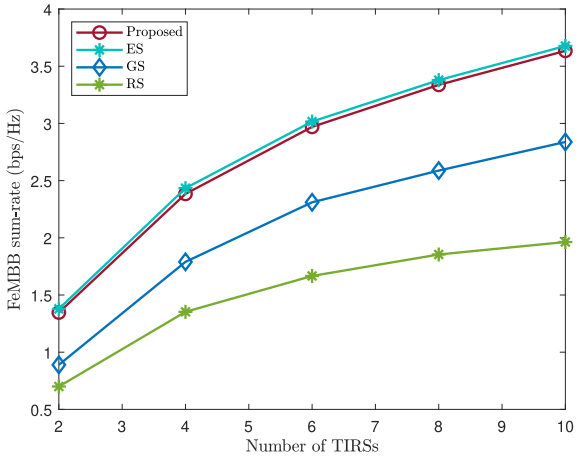


FIGURE 7. FeMBB sum-rate vs. number of TIRSs.

sub-IRSs, the proposed scheme achieves FeMBB sum rates, which is approximately 25% and 46% higher than those of the GS and RS schemes, respectively.

Fig. 6 clearly shows that the sum rate of FeMBB UEs increases with the number of antennas at the AP. This improvement is due to the proportional rise in total signal gain as the number of antennas increases, resulting in higher sum rates. The proposed scheme consistently outperforms both the GS and RS schemes. Additionally, the sum rate of FeMBB UEs achieved with the proposed algorithm is approximately equal to that obtained with the ES scheme.

With an increase of the sum rate of FeMBB UEs, the number of TIRSs increases as shown in Fig. 7. The reason behind this is that when the number of TIRSs increases, the total received signal strength also increases, thereby ultimately improving the sum rate of FeMBB UEs.Furthermore, the sum rate of the proposed scheme is comparable to that of the ES scheme. Fig. 8 shows that as the ratio of eURLLC UEs changes, the proposed method results in lower FeMBB rate loss compared to the GS and RS methods. This improvement is due to the proposed matching-based approach, where each

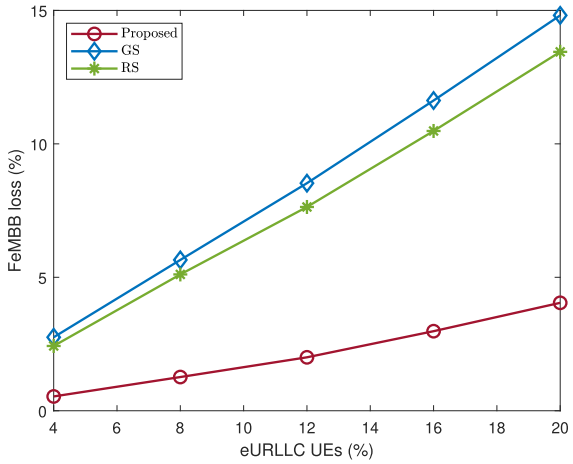


FIGURE 8. FeMBB rate loss (%) vs. eURLLC UEs (%).

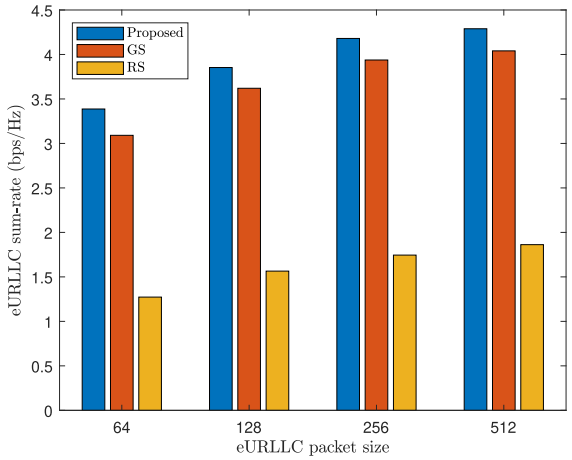


FIGURE 10. eURLLC sum-rate vs. the eURLLC packet size.

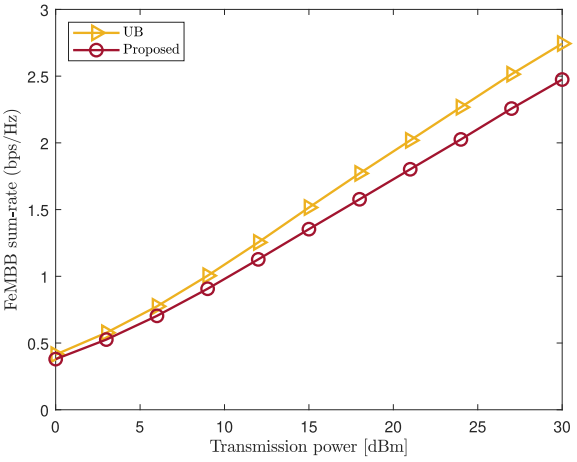


FIGURE 9. FeMBB sum-rate vs. transmit power for UB.

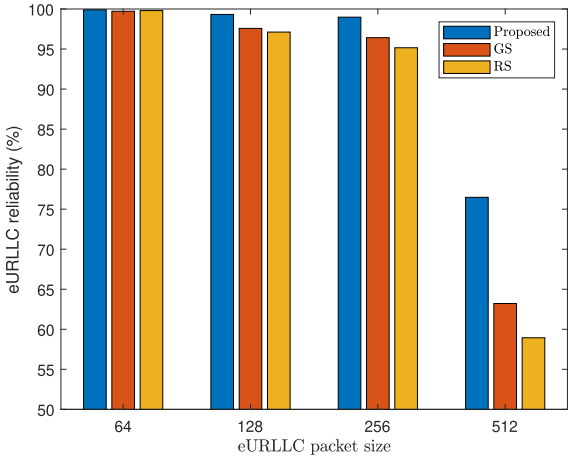


FIGURE 11. eURLLC reliability vs. the eURLLC packet size.

eURLLC UE is paired with the sub-IRS that has the least impact on the FeMBB UEs, which enables the achievement of less FeMBB rate loss.

Fig. 9 shows the FeMBB UEs sum rate as a function of transmit power for the upper bound (UB), representing the maximum when only FeMBB users are considered in the network, and the proposed scheme, which optimizes FeMBB service with the coexistence of eURLLC service. The results demonstrate that the proposed scheme achieves a sum-rate performance near the UB, showcasing its efficiency in maintaining the FeMBB UEs sum rate despite additional constraints introduced by supporting eURLLC UEs. This outcome highlights effectiveness of the proposed algorithm in balancing the requirements of both services while ensuring minimal degradation in FeMBB performance.

B. PERFORMANCE EVALUATION OF eURLLC

Fig. 10 shows the impact of eURLLC packet size on the eURLLC sum rate for various schemes. The sum rate increases with the packet size as more channel resources can be utilized for transmission. Moreover, the proposed matching-based method achieves a higher sum rate for eURLLC UEs compared to the GS and RS schemes.

Fig. 11 illustrates the impact of the eURLLC packet size on the eURLLC reliability of various schemes. The proposed algorithm outperforms the GS and RS methods. Additionally, Fig. 11 shows that as the size of eURLLC packets increases, eURLLC reliability decreases. This decline is due to limited resources, which result in more eURLLC packets being dropped. Fig. 12 provides an illustration of the impact of the number of eURLLC UEs on the eURLLC reliability of different schemes. It shows that as the number of eURLLC UEs increases, the proposed scheme maintains a reliability of 99.9%. However, the reliability of the baseline schemes decreases as the number of eURLLC UEs rises.

Fig. 13 illustrates the eURLLC sum rate of various schemes against the transmit power of the AP. As the transmission power increases, the eURLLC sum rate consistently rises across all schemes. The proposed algorithm, in particular, shows superior performance compared to the GS and RS algorithms. Fig. 14 demonstrates the effect of transmit power

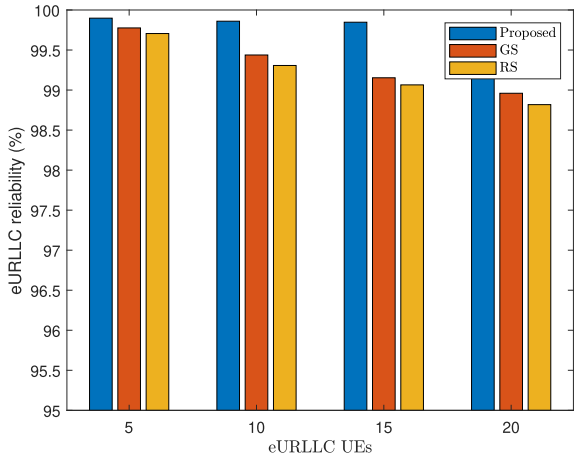


FIGURE 12. eURLLC reliability vs. number of eURLLC UEs for $\zeta = 64$.

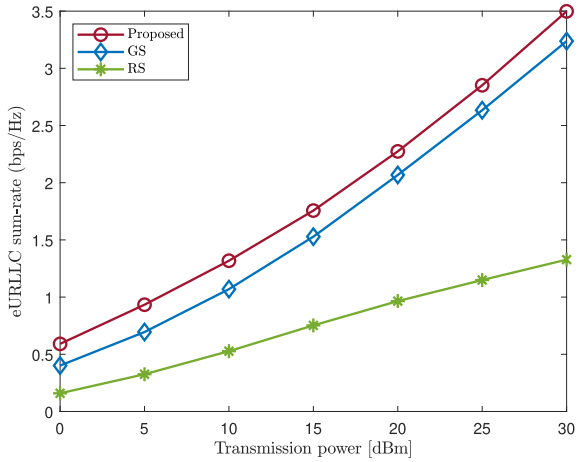


FIGURE 13. eURLLC sum-rate vs. transmit power.

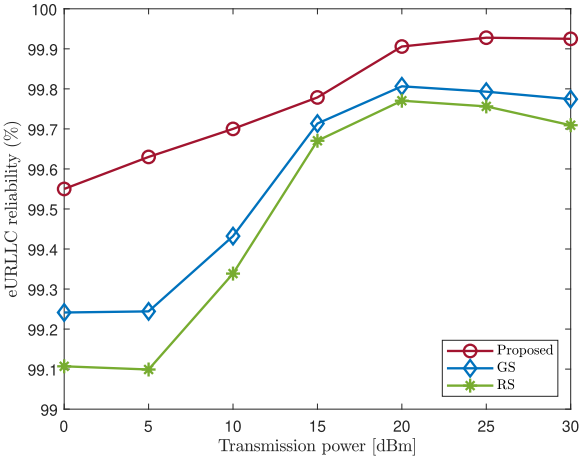


FIGURE 14. eURLLC reliability vs. transmit power.

on eURLLC reliability across various schemes. The figure shows that eURLLC reliability improves with the increase in the transmission power across all schemes. Moreover, the eURLLC reliability sharply increases as the transmit power rises from 5 to 20 dBm. This behavior is primarily due to the significant improvement in the received SINR in this regime, which enhances reliability under the finite blocklength transmission model. However, when the transmit power exceeds 20 dBm, reliability improvement saturates and exhibits a slight degradation. This change can be attributed to diminishing SINR gains and the increased influence of residual interference and noise amplification, which become dominant at higher power levels. Furthermore, the proposed algorithm achieves the highest eURLLC reliability among all schemes, as shown in Fig. 14.

VI. CONCLUSION

In conclusion, this paper examined the coexistence of FeMBB and eURLLC services within an IRS-assisted hybrid mmWave and THz network. To mitigate the obstruction of LOS communication, we proposed a hybrid deployment scheme that combines TIRS and AIRS. Additionally, we introduced a sub-IRS concept, where the IRS is divided into multiple sub-IRSs to serve multiple users. The coexistence of FeMBB and eURLLC services in the same network posed a significant RA challenge. To address this challenge, we formulated a multi-objective optimization problem that jointly optimizes power, user, and service allocation. Our objective was to maximize the FeMBB data rate and eURLLC reliability, which constitutes a challenging NP-hard mixed-integer nonlinear programming problem. To solve this, we utilized a weighted sum method to convert the problem into a single-objective optimization problem, decomposing it into FeMBB and eURLLC RA sub-problems. Specifically, we introduced a many-to-many matching game to allocate IRSs to FeMBB UEs. Simulation results revealed the superior performance of the proposed scheme over baseline methods, particularly in terms of the sum data rate for FeMBB UEs and the reliability of eURLLC UEs. Furthermore, the proposed algorithm's sum rate for FeMBB UEs closely approximates that of the ES method, but with significantly lower computational complexity. These findings underscore the effectiveness of the proposed HIRS deployment scheme and the sub-IRS concept in enhancing the performance of future 6G networks. In future studies, we plan on incorporating advanced beamforming techniques, such as hybrid beamforming, and considering scenarios with multiple antennas at the user equipment. In addition, a promising direction for further investigation is extending the proposed framework to advanced IRS architectures, such as beyond-diagonal IRS (BD-IRS) and simultaneously transmitting and reflecting IRS (STAR-IRS), as well as incorporating hardware-aware modeling that accounts for mutual coupling, radiation loss, and synchronization imperfections Another direction for further research aimed at addressing highly non-convex resource allocation problems in large-scale RIS-assisted networks is using machine learning (ML)-based approaches. In particular, deep reinforcement learning and data-driven optimization techniques may reduce

the need for explicit convexification and manual subproblem decomposition.

APPENDIX A
PROOF OF THEOREM 1

The Lagrangian function associated with **P2.2** can be expressed as

$$\begin{aligned}
L(\mathbf{P}^e, \mu^e, \varrho^e, \upsilon^e) &= \sum_{i=1}^{I} R_i^e - \mu^e\big(\sum_{i=1}^{I} p_i - P_{\max}^e\big) \\
&\quad - \varrho_i^e p_i - \sum_{i=1}^{I} \upsilon_i^e \big(R_{\min} - R_{l,c,i}^e\big) \\
&= \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i} R_{l,c,i}^e - \mu^e\big(\sum_{i=1}^{I} p_i - P_{\max}^e\big) \\
&\quad - \varrho_i^e p_i - \sum_{i=1}^{I} \upsilon_i^e \big(R_{\min} - R_{l,c,i}^e\big) \\
&= \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C}(1+\upsilon_i^e)[\Psi]_{(l,c),i} R_{l,c,i}^e - \mu^e\big(\sum_{i=1}^{I} p_i - P_{\max}^e\big) \\
&\quad - \varrho_i^e p_i - \sum_{i=1}^{I} \upsilon_i^e R_{\min}, \qquad (41)
\end{aligned}$$

where μ^e, ϱ_i^e, and υ_i^e are the Lagrangian multipliers for the total power constraint, the non-negativity constraints on p_i, and the minimum rate constraints, respectively. Substituting (20) in (41), we have

$$\begin{aligned}
L(\mathbf{P}^e, \mu^e, \varrho^e, \upsilon^e) &= \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C}(1+\upsilon_i^e)[\Psi]_{(l,c),i} \\
&\quad \times \log_2\Big(1 + \frac{p_i |(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_i|^2}{\iota_i(t) + \sigma_i^2}\Big)\Big) - \varrho_i^e p_i \\
&\quad - \mu^e\big(\sum_{i=1}^{I} p_i - P_{\max}^e - \sum_{i=1}^{I} \upsilon_i^e R_{\min}, \qquad (42)
\end{aligned}$$

where $\iota_i(t) = \sum_{i'=1 i'\neq i}^{I}\sum_{l=1}^{L}\sum_{c=1}^{C} p_{i'} |(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_{i'}|^2$ is the interference term. By taking the derivative of (42) with respect to $p_{c,i}$ and setting it to zero, we can derive the Karush-Kuhn-Tucker (KKT) condition for the FeMBB UEs' power optimization problem as

$$\begin{aligned}
&\frac{\partial L(\mathbf{P}^e, \mu^e, \varrho^e, \upsilon^e)}{\partial p_{c,i}} = 0 \\
&\frac{(1+\upsilon_i^e)\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i}|(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}{\iota_i(t)+\sigma_i^2+\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i} p_i |(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_{c,i}|^2} - \mu^e - \varrho_i^e = 0 \\
&\frac{\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i}|(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}{\iota_i(t)+\sigma_i^2+\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i} p_i |(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_{c,i}|^2} = \frac{\mu^e + \varrho_i^e}{1+\upsilon_i^e} \\
&p_i + \frac{\iota_i(t)+\sigma_i^2}{\sum_{l=1}^{L}\sum_{c=1}^{C}[\Psi]_{(l,c),i}|(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2} = \frac{1+\upsilon_i^e}{\mu^e + \varrho_i^e}. \qquad (43)
\end{aligned}$$

We rearrange (43) to calculate p_i, which can be written as

$$p_i = \Big[\frac{1+\upsilon_i^e}{\mu^e + \varrho_i^e} - \frac{\iota_i(t)+\sigma_i^2}{\sum_{l\in L}\sum_{c\in C}[\Psi]_{(l,c),i}|(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}\Big]^+, \qquad (44)$$

where $[x]^+ = \max\{x, 0\}$. Furthermore, to find eMBB UEs' power budget, we substitute (44) into (28d), yielding

$$P_{\max}^e = \sum_{i=1}^{I}\Big[\frac{1+\upsilon_i^e}{\mu^e + \varrho_i^e} - \frac{\iota_i(t)+\sigma_i^2}{\sum_{l\in L}\sum_{c\in C}[\Psi]_{(l,c),i}|(\mathbf{h}_{l,c,i}^{\mathbb{T}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_i|^2}\Big]^+. \qquad (45)$$

This completes the proof of Theorem 1.

APPENDIX B
PROOF OF THEOREM 2

The Lagrangian function for **P3.2** can be expressed as

$$\begin{aligned}
L(\mathbf{P}^u, \mu^u, \varrho_i^u) &= \sum_{j=1}^{J}\sqrt{\frac{\zeta_j}{V_j}}\big(\log_2(1+\gamma_{l,c,j}^u) - R_{l,c,j}^u\big) \\
&\quad - \mu^u\big(\textstyle\sum_{j=1}^{J} p_j - P_{\max}^u\big) - \varrho_i^u p_j, \\
&= \sum_{j=1}^{J}\sqrt{\frac{\zeta_j}{V_j}}\Big(\log_2\big(1 + \frac{p_j|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}{\iota_j(t)+\sigma_j^2}\big) - R_{l,c,j}^u\Big) \\
&\quad - \mu^u\big(\textstyle\sum_{j=1}^{J} p_j - P_{\max}^u\big) - \varrho_i^u p_j, \qquad (46)
\end{aligned}$$

where μ^u and ϱ_i^u are the Lagrangian multipliers for the total power constraint and the non-negativity constraints on p_j, respectively, and $\iota_j(t) = \sum_{j'=1,j'\neq j}^{J}\sum_{l=1}^{L}\sum_{c=1}^{C} p_{j'}|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_{j'}|^2$ is the interference term. By taking the derivative of (46) with respect to p_j and setting it to zero, we can derive the KKT condition for the eURLLC UEs' power optimization problem as

$$\begin{aligned}
&\frac{\partial L(\mathbf{P}^u, \mu^u, \varrho_i^u)}{\partial p_j} = 0 \\
&\sqrt{\frac{\zeta_j}{V_j}}\frac{|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}{\iota_j(t)+\sigma_j^2+p_j|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2} - \mu^u - \varrho_i^u = 0 \\
&p_j + \frac{\iota_j(t)+\sigma_j^2}{|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2} = \sqrt{\frac{\zeta_j}{V_j}}\big(\frac{1}{\mu^u + \varrho_i^u}\big). \qquad (47)
\end{aligned}$$

We rearrange (47) to calculate p_j, which can be written as

$$p_j = \Big[\sqrt{\frac{\zeta_j}{V_j}}\big(\frac{1}{\mu^u + \varrho_i^u}\big) - \frac{\iota_j(t)+\sigma_j^2}{|(\mathbf{h}_{l,c,j}^{\mathbb{M}}(t))^{\mathsf{H}}\widehat{\mathbf{w}}_j|^2}\Big]^+. \qquad (48)$$

Furthermore, to find the URLLC UEs' power budget, we substitute (48) into (28i), yielding

$$P^{u}_{\max} = \sum_{j=1}^{J} \left[\sqrt{\frac{\zeta_j}{V_j}} \left(\frac{1}{\mu^u + \varrho^u_i} \right) - \frac{\iota_j(t) + \sigma_j^2}{|(\mathbf{h}^{\mathbb{M}}_{l,c,j}(t))^{\mathsf{H}} \widehat{\mathbf{w}}_j|^2} \right]^{+}. \quad (49)$$

This completes the proof of Theorem 2.

REFERENCES

[1] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[3] L. Zhang, Y.-C. Liang, and D. Niyato, "6G visions: Mobile ultra-broadband, super Internet-of-Things, and artificial intelligence," *China Commun.*, vol. 16, no. 8, pp. 1–14, Aug. 2019.

[4] J. Park et al., "Extreme URLLC: Vision, challenges, and key enablers," 2020, *arXiv:2001.09683*.

[5] Y. Guo, Z. Qin, Y. Liu, and N. Al-Dhahir, "Intelligent reflecting surface aided multiple access over fading channels," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 2015–2027, Mar. 2021.

[6] M. Rahim, A. Basit, B. Selim, and G. Kaddoum, "Hierarchical IRS-assisted user association for sum-rate maximization in hybrid THz LEO-HAPS-terrestrial networks," *IEEE Commun. Lett.*, vol. 29, no. 12, pp. 2919–2923, Dec. 2025.

[7] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.

[8] Q. Chen, J. Wu, J. Wang, and H. Jiang, "Coexistence of URLLC and eMBB services in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 839–851, Jan. 2023.

[9] M. Rahim, T. Luan Nguyen, T. Nhu Do, and G. Kaddoum, "Joint power and user allocation in coexistence of eMBB and URLLC services," *IEEE Commun. Lett.*, vol. 28, no. 9, pp. 2186–2190, Sep. 2024.

[10] M. Rahim, T. Luan Nguyen, and G. Kaddoum, "JPUSA in coexistence of eMBB and URLLC services in multi-cell IRS-assisted terahertz networks," *IEEE Trans. Green Commun. Netw.*, vol. 9, no. 3, pp. 1206–1223, Sep. 2025.

[11] T. S. Rappaport et al., "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[12] *IEEE Standard for High Data Rate Wireless Multi-media Networks-Amendment 2: 100 Gb/s Wireless Switched Point-to-Point Physical Layer*, Standard IEEE 802, 2017, pp. 1–55.

[13] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[14] M. Di Renzo et al., "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 798–807, 2020.

[15] T. D. Hua, M. Mohammadi, H. Q. Ngo, and M. Matthaiou, "Cell-free massive MIMO SWIPT with beyond diagonal reconfigurable intelligent surfaces," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2024, pp. 1–6.

[16] T. D. Hua, M. Mohammadi, H. Q. Ngo, and M. Matthaiou, "Cell-free massive MIMO-assisted SWIPT using stacked intelligent metasurfaces," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 13056–13072, 2026.

[17] D.-T. Hua, Q. T. Do, N.-N. Dao, T.-V. Nguyen, D. Shumeye Lakew, and S. Cho, "Learning-based reconfigurable-intelligent-surface-aided rate-splitting multiple access networks," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17603–17619, Oct. 2023.

[18] M. Rahim, T. L. Nguyen, G. Kaddoum, and T. N. Do, "Multi-IRS-aided terahertz networks: Channel modeling and user association with imperfect CSI," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 836–855, 2024.

[19] M. Rahim, G. Kaddoum, and T. N. Do, "Joint devices and IRSs association for terahertz communications in industrial IoT networks," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 375–390, Mar. 2024.

[20] J. Li and X. Zhang, "Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543–1546, Sep. 2020.

[21] Y. Ruan, G. Nie, W. Ni, H. Tian, and J. Ren, "Efficient traffic scheduling for coexistence of eMBB and uRLLC in industrial IoT networks," in *Proc. IEEE WCNC*, 2022, pp. 1431–1436.

[22] T. Pan, X. Wu, T. Zhang, and X. Li, "Energy-efficient resource allocation in ultra-dense networks with EMBB and URLLC users coexistence," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 2549–2563, Feb. 2024.

[23] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint scheduling of eMBB and URLLC services in RIS-aided downlink cellular networks," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Athens, Greece, Jul. 2021, pp. 1–9.

[24] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, Feb. 2022.

[25] H. Zarini, N. Gholipoor, M. R. Mili, M. Rasti, H. Tabassum, and E. Hossain, "Resource management for multiplexing eMBB and URLLC services over RIS-aided THz communication," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1207–1225, Feb. 2023.

[26] M. Rahim and S. Cherkaoui, "Dual-tier IRS-assisted mid-band 6G mobile networks: Robust beamforming and user association," 2026, *arXiv:2602.00431*.

[27] M. Rahim and S. Cherkaoui, "Reliable IoT communications in 6G non-terrestrial networks with dual RIS," 2026, *arXiv:2602.00438*.

[28] K. Humadi, I. Trigui, W.-P. Zhu, and W. Ajib, "User-centric cluster design and analysis for hybrid sub-6GHz-mmWave-THz dense networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7585–7598, Jul. 2022.

[29] P. U. Adamu, M. López-Benítez, and J. Zhang, "Hybrid transmission scheme for improving link reliability in mmWave URLLC communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6329–6340, Sep. 2023.

[30] Z. Tian, J. Wang, Z. Chen, M. Wang, Y. Jia, and D. O. Wu, "Hybrid deployment of multi-hierarchical RISs for probabilistic LoS communication in dense urban," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 2207–2211, Aug. 2023.

[31] M. Rasti, S. K. Taskou, H. Tabassum, and E. Hossain, "Evolution toward 6G multi-band wireless networks: A resource management perspective," *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 118–125, Aug. 2022.

[32] Z. Wan, Z. Gao, F. Gao, M. D. Renzo, and M.-S. Alouini, "Terahertz massive MIMO with holographic reconfigurable intelligent surfaces," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4732–4750, Jul. 2021.

[33] Y. Fang, S. Atapattu, H. Inaltekin, and J. Evans, "Optimum reconfigurable intelligent surface selection for wireless networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6241–6258, Sep. 2022.

[34] M. Haghshenas, P. Ramezani, M. Magarini, and E. Björnson, "Parametric channel estimation with short pilots in RIS-assisted near- and far-field communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 10366–10382, Aug. 2024.

[35] S. Kim and M. Kim, "Rotation representations and their conversions," *IEEE Access*, vol. 11, pp. 6682–6699, 2023.

[36] B. Al-Nahhas, Q. Nadeem, and A. Chaaban, "Intelligent reflecting surface assisted MISO downlink: Channel estimation and asymptotic analysis," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[37] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

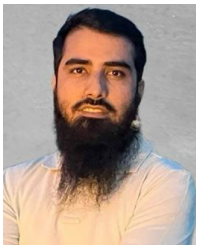
[38] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Superposition-based URLLC traffic scheduling in 5G and beyond wireless networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6295–6309, Sep. 2022.

[39] A. Ranjha and G. Kaddoum, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, Mar. 2021.

[40] A. Ranjha, G. Kaddoum, M. Rahim, and K. Dev, "URLLC in UAV-enabled multicasting systems: A dual time and energy minimization problem using UAV speed, altitude and beamwidth," *Comput. Commun.*, vol. 187, pp. 125–133, Apr. 2022.

[41] K. Dovelos, S. D. Assimonis, H. Quoc Ngo, B. Bellalta, and M. Matthaiou, "Intelligent reflecting surfaces at terahertz bands: Channel modeling and analysis," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[42] H. Shakhatreh, A. Sawalmeh, A. H. Alenezi, S. Abdel-Razeq, M. Almutiry, and A. Al-Fuqaha, "Mobile-IRS assisted next generation UAV communication networks," 2022, *arXiv:2207.03622*.
[43] H. Zhang, Y. Duan, K. Long, and V. C. M. Leung, "Energy efficient resource allocation in terahertz downlink NOMA systems," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1375–1384, Feb. 2021.
[44] R. Liu, G. Yu, J. Yuan, and G. Y. Li, "Resource management for millimeter-wave ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1094–1108, Feb. 2021.
[45] Y. Xu and S. Mao, "User association in massive MIMO HetNets," *IEEE Syst. J.*, vol. 11, no. 1, pp. 7–19, Mar. 2017.
[46] S. Sun, F. Yang, and J. Song, "Sum rate maximization for intelligent reflecting surface-aided visible light communications," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3619–3623, Nov. 2021.
[47] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.
[48] S. Nie, G. R. MacCartney, S. Sun, and T. S. Rappaport, "28 GHz and 73 GHz signal outage study for millimeter wave cellular and backhaul communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 4856–4861.



MUDDASIR RAHIM received the B.S. and M.S. degrees in electrical (telecommunication) engineering from COMSATS University Islamabad, Pakistan, in March 2016 and September 2020, respectively, and the Ph.D. degree (Hons.) from the École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, Canada, in 2025. He was a Research Assistant with the Department of Electrical and Computer Engineering, COMSATS University Islamabad, from 2019 to 2020. His research interests include cognitive radio networks, device-to-device communication, wireless sensor networks, intelligent reconfigurable surface, matching theory, and terahertz communications.

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the École Nationale Supérieure de Techniques Avancées (ENSTA Bretagne), Brest, France, the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the Université de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences (INSA), University of Toulouse, Toulouse, France, in 2009. He is currently a Professor and the Research Director of the Resilient Machine Learning Institute (ReMI), and the Tier 2 Canada Research Chair of the École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, Canada. He has published over more than 300 journals, conference papers, two chapters in books, and has eight pending patents. His recent research interests include wireless communication networks, tactical communications, resource allocations, and network security. He received the Best Papers Awards at the 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications (WIMOB), the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), and the 2023 IEEE International Wireless Communications and Mobile Computing Conference (IWCMC). Moreover, he received IEEE TRANSACTIONS ON COMMUNICATIONS Exemplary Reviewer Award in 2015, 2017, and 2019. In addition, he received the Research Excellence Award of the Université du Québec in 2018. In 2019, he received the Research Excellence Award from ÉTS in recognition of his outstanding research outcomes. He also received the 2022 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence (Middle Career Researcher). Lastly, he received the prestigious 2023 MITACS Award for Exceptional Leadership. He served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE COMMUNICATIONS LETTERS. He serving as an Area Editor for IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING and an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.