Active RIS in Digital Twin-based URLLC IoT Networks: Fully-Connected vs. Sub-Connected?

Sravani Kurma *Graduate Student Member, IEEE*, Tri Ayu Lestari *Graduate Student Member, IEEE*, Keshav Singh, *Member, IEEE*, Anal Paul *Member, IEEE*, and Shahid Mumtaz, *Senior Member, IEEE*

***Abstract*—The substantial power consumption attributed to the active components within fully-connected reconfigurable intelligent surface (RIS) architecture significantly hinders the efficiency and sustainability of DT-enabled MEC networks. To tackle this challenge, we present an innovative sub-connected architecture for active RIS within the digital twin (DT) integrated mobile edge computing (MEC) framework of an Internet-of-Things (IoT) networks, capitalizing on edge intelligence to enhance ultra-reliable and low-latency communication (URLLC) services. The primary aim of our research is to improve uplink data transmission from IoT URLLC user nodes (UNs) to a base station (BS) with the aid of an active RIS, even under an imperfect channel state information (CSI). We have formulated the total end-to-end (e2e) latency minimization problem, which is solved by using an efficient alternating optimization (AO) algorithm. The algorithm breaks down the proposed non-convex problem into five subproblems, namely, beamforming design, caching and offloading policy optimization, joint communication and computation optimization, and joint active RIS phase shift and amplification factor vector optimization. We conducted a thorough analysis of the convergence properties of the proposed AO algorithm, benchmarking its performance against the established Heuristic algorithm. Our simulation results consistently demonstrate the superiority of our proposed DT-assisted optimal phase sub-connected active RIS scheme over various benchmark schemes, taking into account various factors such as the number of RIS elements, power budget constraints, imperfect CSI , edge computing server (ECS) cache capacity, number of IoT UNs, and the number of power amplifiers.**

***Index Terms*—Active reconfigurable intelligent surface, fully-connected architecture, sub-connected architecture, ultra-reliable and low latency communication, digital twin, alternating optimization, mobile edge computing.**

I. Introduction

INTERNET-OF-THINGS (IoT) connectivity is vital for time-critical communication, ensuring data delivery within specified latency bounds. In the realm of advanced wireless communication services, particularly ultra-reliable and low-latency communication (URLLC), the demand for massive IoT user nodes (UNs) connectivity and latency-sensitive applications poses significant challenges [1]. The remarkable technology of mobile edge computing (MEC) empowers URLLC-based industrial IoT applications to achieve exceptional services and experiences [2]. Achieving efficient task offloading in edge computing environments is challenged by many factors, such as joint computations, heterogeneous architecture, and task integration. To address these challenges and enhance task offloading effectiveness, the integration of digital twin (DT) and Metaverse technologies has emerged as a promising approach [3], [4]. DT represents a virtual replica of a physical object, system, or process, enabling simulation, analysis, and optimization. Metaverse, on the other hand, is a virtual environment that allows physical objects (e.g., IoT UNs) to interact with a virtual or digital environment [5]. The integration of DT and MEC offers seamless end-to-end (e2e) Metaverse services using real-time optimization theory [6], [7]. It should be noted that most existing works assume a direct link to the MEC, which is quite impractical in many scenarios due to obstacles obstructing a direct link. To solve this problem, alternative transmission paths can be provided by the reconfigurable intelligent surface (RIS) [8]. While the passive RIS relies solely on phase-shift control for signal reflection, the active RIS introduces a game-changing twist by integrating individual power amplifiers within each RIS element, thus enabling active signal amplification [9]. Consequently, harnessing the strengths of active RIS, equipped with integrated reflection-type amplifiers to enhance received signals, holds the potential to substantially boost the data rate for task offloading to the MEC [10], [11]. However, the presence of dedicated power amplifiers for each programmable element in active RIS can pose a substantial challenge in terms of high power consumption [12], [13]. Hence, a sub-connected architecture is explored to address this issue. In a sub-connected setup, power amplifiers are shared among groups of RIS elements, improving energy efficiency while optimizing signal paths and reducing interference, thus enhancing signal quality [8], [14], [15]. This results in faster data transmission and reduced latency, enabling signals to travel more efficiently through optimized pathways. The synergy between the Metaverse, DT networks, and sub-connected active RIS is paving the way for communication technology advancements that offer reliable communication while ensuring minimal delays for vital IoT applications, such as industrial IoT automation and IoT-enabled autonomous vehicles.

The work of K. Singh was supported by the National Science and Technology Council, Taiwan under Grant NSTC 112-2923-E-008-003-MY3, while the work of S. Mumtaz was supported by the 6G-SENSES project from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101139282. (*Corresponding author: Keshav Singh*)

S. Kurma, T. A. Lestari, K. Singh, and A. Paul are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan. (E-mail: sravani.phd.nsysu.21@gmail.com, triayulestari19@gmail.com, keshav.singh@mail.nsysu.edu.tw, apaul@ieee.org).

S. Mumtaz is with the Department of Applied Informatics, Silesian University of Technology Akademicka 16 44-100 Gliwice, Poland and Nottingham Trent University, Department of Computer Sciences, Nottingham NG14FQ, U.K. (E-mail: dr.shahid.mumtaz@ieee.org).

A. Related works

Previous studies [16], [17] primarily focused on resource allocation for secure URLLC, with an emphasis on communication aspects at the expense of computation. However, given the demand for mission-critical application devices to execute computation-intensive tasks within tight time constraints, MEC has become a compelling solution for enabling rapid and efficient computation in URLLC systems [10], [18], [19]. Simultaneously, DT technology offers viable features that enable organizations to adhere to the strict needs of high reliability and low latency, ensuring uninterrupted and dependable operations [1], [10]. In the context of IoT networks, where seamless and reliable connectivity is paramount, the integration of DT with URLLC becomes especially important, as it facilitates the optimization of configurations to meet URLLC requirements [4], [10], [19].

In the ever-expanding realm of IoT networks within bustling urban environments, the quest for enhanced connectivity has given rise to a wave of innovative research on RIS [12], [20], [21], [21]–[23]. In particular, recent research works show the efficacy of active RIS over passive RIS [12], [13]. However, prior research has primarily concentrated on active RIS systems employing a dedicated power amplifier for each individual RIS element, a setup commonly referred to as the fully-connected architecture [12], [13]. This setup can pose a significant power consumption challenge, especially with a large number of RIS elements, which is addressed through the use of a sub-connected architecture [14], [15], [24]–[26]. Prior research predominantly centered on various aspects of beamforming, including joint beamforming design for passive RIS [24], active RIS [15], hybrid RIS [24], [25], as well as channel estimation [26], all within the context of the infinite blocklength regime. Motivated by the true potential of RIS technology, unquestionably, there is a significant research scope in RIS-assisted URLLC services to enhance the seamless experiences of delay-sensitive UN [22], [27]–[29]. However, only a limited number of studies have specifically explored the application of RIS in URLLC scenarios [30]–[32]. The authors in [30] proposed a joint optimization of phase shifts and beamforming variables of an active RIS to allocate URLLC traffic, aiming to maximize the URLLC sum rate in a multiple-input single-output system, where a group of base station (BS) collaborates to serve URLLC traffic. In [10], [33], the authors proposed an RIS-assisted DT-enabled URLLC service to enhance reliability and reduce the transmission delay while offloading the task to the BS from the UN. Table I offers an extensive investigation of the existing works and highlights the contributions of our proposed work.

B. Motivations and Contributions of the Work

The motivation behind this research is multi-faceted. Firstly, in addressing the challenges posed by the power consumption in the fully-connected RIS architecture [12] within RIS-aided IoT networks, there is a need to incorporate a sub-connected active RIS architecture [14]. Secondly, there is a crucial requirement to advance the technological frontier by harnessing the potential of the edge intelligence to unlock new capabilities in URLLC services in MEC-enabled DT networks [1], [4], [34]. Addressing these requirements is essential to empower industries and applications that depend on real-time, mission-critical communication and foster the growth of IoT-enabled innovations. Our study stands at the crossroads in this pursuit, poised to unravel the trade-offs and advantages between fully-connected and sub-connected RIS architectures in DT-based MEC communication frameworks, paving the way for future research and advancements in this emerging field. The key contributions of our proposed work can be summed up as follows:

- We investigate a DT-based MEC system assisted by a fully-connected and sub-connected active RIS architecture to facilitate task offloading and enhance IoT-URLLC services. The main aim of our work is to minimize the total e2e latency in the proposed system while considering various constraints, such as beamforming, edge caching, transmit power, task-offloading policies, energy consumption, processing rates of the IoT UN, edge computing server (ECS), active-RIS phase shift matrices, amplification factor vector, and allocated bandwidth for each IoT UN.
- We develop an efficient algorithm to address the problem by breaking it down into five subproblems: beamforming design, joint communication and computation optimization, offloading policy optimization, caching policy optimization, and joint active RIS phase shift and amplification factor vector optimization.
- In our study, we conduct simulations to compare latency convergence in fully-connected versus sub-connected configurations using the AO algorithm, benchmarked against a Heuristic algorithm. These comparisons were based on multiple parameters, including the number of RIS elements, power budget constraints, imperfect CSI, ECS cache capacity, the number of IoT UNs, and the number of power amplifiers. The findings consistently indicate a preference for the sub-connected active RIS configuration. This configuration demonstrates the superior performance in latency reduction, attributed to optimized signal paths, minimized interference, and improved signal quality, thereby facilitating faster data transmission.

The remainder of the paper is organized as follows: Section II describes the proposed system model. Section III presents the proposed solutions for fully-connected active RIS, while Section IV presents the solutions for the sub-connected active RIS. An extensive numerical analysis is used to demonstrate the efficacy of the considered network is discussed in Section V. Finally, in Section VI, our proposed work is concluded.
Notations: For the reader's convenience and clarity of understanding, all essential symbols, along with their definition, are comprehensively outlined in Table II.

II. System Model

As depicted in Fig. 1 and Fig. 2, we examine the fully-connected and sub-connected active RIS-assisted MEC-enabled DT networks, respectively. We consider the randomly distributed K IoT UN within a specific urban setting, denoted

TABLE I: A Comparative overview of our work and the state-of-the-art

Paper	RIS	RIS architecture	MEC	URLLC	Algorithm	DT	Performance metric
[1]	✗	✗	✓	✓	AO	✓	Worst-case latency minimization of e2e DT latency
[2]	✗	✗	✓	✗	AO	✗	Aggregative game-based task partitioning and offloading scheme
[4]	✗	✗	✓	✓	AO	✓	Total e2e latency minimization
[8]	Passive	FC	✗	✓	AO	✗	Average decoding error probability and data rate
[10]	Passive	FC	✓	✓	DRL	✓	Latency minimization
[12]	Active, Passive	FC	✗	✗	AO	✗	Sum-rate maximization
[13]	Active, Passive	FC	✗	✗	AO	✗	Rate maximization
[14]	Active	SC	✗	✗	AO	✗	Energy efficiency maximization
[15]	Active	SC	✗	✓	AO	✗	Sum-rate maximization and power minimization
[16]	✗	✗	✗	✓	AO	✗	Latency minimization
[17]	✗	✗	✗	✓	SCA	✗	Minimization of the total transmit power
[19]	✗	✗	✓	✗	AO	✓	Latency minimization
[20]	Passive	FC	✗	✗	AO	✓	Rate maximization
[22]	Passive	FC	✗	✗	Dinkelbach's Method	✗	Energy efficiency maximization
[33]	Passive	FC	✗	✓	SGD, MO-SAC	✗	Transmission latency and the total service cost minimization
[34]	✗	✗	✓	✓	DDQN	✓	Energy consumption minimization
[35]	✗	✗	✓	✓	DDN	✓	Energy consumption minimization
[36]	✗	✗	✓	✗	PPO	✓	Energy consumption minimization
Our work	Active	FC and SC	✓	✓	AO	✓	Total e2e latency minimization

by $\mathcal{K} = \{1,2,...,K\}$. For the purpose of task offloading, single-antenna IoT UNs establish communication with a M-antenna BS, operating within the constraints of finite block-length transmission. We assume the scenario where the direct path is obstructed in the urban environment due to the dense network coverage. Hence, we employ an active RIS[1] comprising N reflecting elements, represented by the set $\mathcal{N} = \{1, 2, ..., N\}$. Every reflecting element of the active RIS introduces an additional reflection, characterized by a reflection coefficient $\phi_n \triangleq \alpha_n e^{j\varrho_n}$, where ϱ_n represents the phase shift. The amplitude of the n^{th} RIS element, denoted as α_n, is subject to the constraint of a maximum amplification gain[2], denoted as α_{max}, such that $\alpha_n \leq \alpha_{max}$. The active RIS phase shift matrix is represented as $\mathbf{\Phi} \triangleq \text{diag}\{\phi\}$, with $\phi \triangleq \{\phi_1, \ldots, \phi_N\}^T \in \mathbb{C}^{N\times N}$.

The channel gain between the IoT UN and the active RIS is denoted by $\mathbf{h}_{r,k} \in \mathbb{C}^{N\times 1}$ and the channel gain between active RIS and BS is represented by $\mathbf{H}_{b,r} \in \mathbb{C}^{M\times N}$ that are defined as follows [37]

$$\mathbf{h}_{r,k} = \frac{d_{r,k}^{\frac{-\gamma_{r,k}}{2}}}{\sqrt{Z_{r,k}+1}}\left(\sqrt{Z_{r,k}}\mathbf{q}_{r,k}^{LoS} + \boldsymbol{\mathcal{T}}_{r,R}^{1/2}\mathbf{q}_{r,k}\right), \quad (1)$$

[1]The manuscript underscores the effectiveness of active RIS systems in diverse scenarios, from urban to remote areas, highlighting their superiority in signal management and energy efficiency compared to traditional relays and IAB nodes, and emphasizes their pivotal role in enhancing smart cities, IoT networks, and challenging communication environments.

[2]The amplification gain α_n is capped at a predefined constant to maintain it within a specified range, thus averting scenarios of excessive amplification that could lead to system instability and adversely affect overall performance.

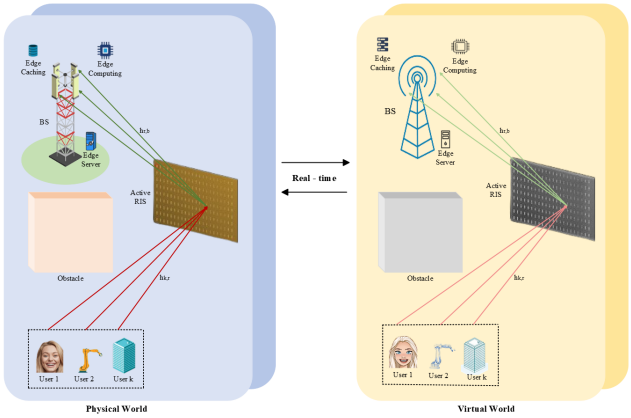

Fig. 1: URLLC fully-connected active RIS-DT system.

$$\mathbf{H}_{r,b} = \frac{d_{b,r}^{\frac{-\gamma_{b,r}}{2}}}{\sqrt{Z_{r,k}+1}}\left(\sqrt{Z_{b,r}}\boldsymbol{\mathcal{Q}}^{LoS} + \boldsymbol{\mathcal{T}}_{b,R}^{1/2}\boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{T}}_{r,T}\right), \quad (2)$$

where $\gamma_{r,k}$ and $\gamma_{b,r}$ are the pathloss exponents. Here, $\mathbf{q}_{r,k}^{LoS}$ and $\boldsymbol{\mathcal{Q}}^{LoS}$ denotes the LoS components from k^{th} IoT UN to the active RIS and from the active RIS to the BS, respectively. Here, $d_{r,k}$ and $d_{b,r}$ are the distance between the IoT UN-RIS and RIS-BS, respectively. The Rician factors are denoted by $Z_{r,k}$ and $Z_{b,r}$ and the small-scale fading parameters are denoted by $\boldsymbol{\mathcal{Q}}$ and $\mathbf{q}_{r,k}$ which follow the standard complex Gaussian distribution. Note that $\boldsymbol{\mathcal{T}}_{r,R}$, $\boldsymbol{\mathcal{T}}_{r,T}$, and $\boldsymbol{\mathcal{T}}_{b,R}$ are the receive correlation matrix at the active RIS, the transmit

TABLE II: Table of Notations

Index		**Meaning**
k	$\in$	$\{1, 2, ..., K\}$ (index of k-th UN)
n	$\in$	$\{1, 2, ..., N\}$ (index of n-th RIS element)
l	$\in$	$\{1, 2, ..., L\}$ (index of l-th required power amplifiers)
Notation		**Meaning**
$\mathbf{h}_{r,k}$	$\triangleq$	channel gain between active RIS and the IoT UN
$\mathbf{H}_{b,r}$	$\triangleq$	channel gain between active RIS and BS
$d_{b,r}$, $d_{r,k}$	$\triangleq$	distance between of the RIS-BS, IoT UN-RIS
$\gamma_{r,k}$, $\gamma_{b,r}$	$\triangleq$	exponents of pathloss
$\mathbf{q}_{r,k}^{LoS}$, $\boldsymbol{\mathcal{Q}}^{LoS}$	$\triangleq$	the LoS components from k^{th} IoT UN to the active RIS, from the active RIS to the BS
$Z_{r,k}$, $Z_{b,r}$	$\triangleq$	the Rician factors
$\boldsymbol{\mathcal{Q}}$, $\mathbf{q}_{r,k}$	$\triangleq$	small-scale fading parameters
$\mathcal{CN}(m, \sigma^2)$	$\triangleq$	Complex Gaussian distribution with mean m and variance σ^2
$\boldsymbol{a}$	$\triangleq$	denotes the amplification factor vector
Γ	$\triangleq$	the indicator matrix showing how the phase-shifting circuits and power amplifiers are connected
L	$\triangleq$	the number of required power amplifiers
γ_k^F	$\triangleq$	SNR of the k^{th} IoT UN for fully-connected
γ_k^S	$\triangleq$	SNR of the k^{th} IoT UN for sub-connected
R_k^F	$\triangleq$	rate of the k^{th} IoT UN for fully-connected
R_k^S	$\triangleq$	rate of the k^{th} IoT UN for sub-connected
δ	$\triangleq$	the transmission time interval
T	$\triangleq$	the cumulative latency
E	$\triangleq$	energy consumption
ξ	$\triangleq$	the energy conversion efficiency
P_{BS}, P_U	$\triangleq$	the dissipated power at BS and each IoT UN
P_{PS}, P_{PA}	$\triangleq$	the active RIS hardware static power
σ_0^2	$\triangleq$	variance of AWGN at active RIS
σ_b^2	$\triangleq$	variance of AWGN at BS
$\lvert\cdot\rvert$	$\triangleq$	absolute value
$\lVert\cdot\rVert$	$\triangleq$	the norm of a value
$Q(\cdot)$	$\triangleq$	complementary cumulative distribution function (CCDF)
$Q^{-1}(\cdot)$	$\triangleq$	the inverse of CCDF
$\mathrm{diag}(\cdot)$	$\triangleq$	diagonal element of a matrix
$Tr(\cdot)$	$\triangleq$	the trace of a square matrix in linear algebra

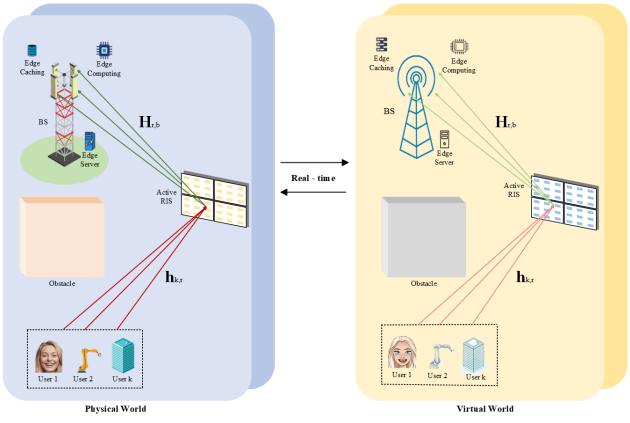

Fig. 2: URLLC sub-connected active RIS-DT system.

correlation matrix at the active RIS, and the receive correlation matrix at the BS, respectively.

By considering an imperfect channel state information (CSI) scenario, the actual channel can be represented in terms of its estimate as $\mathbf{h}_k = \hat{\mathbf{h}}_k + \mathbf{h}_{ek}$, where $\mathbf{h}_k = \mathbf{H}_{r,b}\boldsymbol{\Phi}\mathbf{h}_{r,k}$ denotes the effective channel link from the BS to the IoT UN, $\hat{\mathbf{h}}_k$ is its estimated channel gain, $\mathbf{h}_{ek}$ is the channel estimation error (CEE) vector with each of its entries as independent and identically distributed, following $\mathcal{CN}\left(0, \sigma_{ek}^2\right)$. The signal received at the BS under perfect CSI is given as follows:

$$y_b = \mathbf{w}_k^H(\mathbf{h}_k x_k + \mathbf{H}_{r,b}\boldsymbol{\Phi}\mathbf{z}_0 + \mathbf{n}_b), \forall k \in \mathcal{K}, \tag{3}$$

where $\mathbf{w}_k$ and x_k are active receive beamformer at the BS and the data symbol sent by the k^{th} IoT UN, respectively. Here, $\mathbf{n}_b$ and $\mathbf{z}_0$ represent the additive white Gaussian noise (AWGN) added at the BS and the dynamic noise at the active RIS, which are given as $\mathcal{CN}\left(0, \sigma_b^2 I_M\right)$ and $\mathcal{CN}\left(0, \sigma_0^2 I_N\right)$, respectively. By substituting the $\mathbf{h}_k$ in terms of its estimated value in (3), we get the received signal under imperfect CSI

as

$$y_b = \mathbf{w}_k^H((\hat{\mathbf{h}}_k + \mathbf{h}_{ek})x_k + \mathbf{H}_{r,b}\mathbf{\Phi}\mathbf{z}_0 + \mathbf{n}_b), \forall k \in \mathcal{K}. \quad (4)$$

The following provides the signal-to-noise ratio (SNR) of the k^{th} IoT UN:

$$\gamma_k = \frac{|\mathbf{w}_k^H \hat{\mathbf{h}}_k|^2 p_k}{||\mathbf{w}_k^H \mathbf{H}_{r,b}\mathbf{\Phi}||^2\sigma_0^2 + ||\mathbf{w}_k^H||^2 B b_k N_0}, \quad (5)$$

where B is the total bandwidth of the system and p_k is the transmit power of the k^{th} IoT UN, where $\gamma_k \in \{\gamma_k^F, \gamma_k^S\}$. Here, N_0 represents the effective noise, which combines the noise added at the BS and CEE, i.e., given by $N_0 = \sigma_b^2 + \sigma_{ek}^2$.

A. DT-assisted active RIS communication model

As there is no direct path available due to the obstacles in the path between the IoT UN and the BS, all the IoT UN transmit their signal to the BS with the aid of active RIS. The uplink rate expression for the k^{th} IoT UN is given as [38]

$$R_k = \frac{B}{\ln 2}\left[b_k \ln(1+\gamma_k) - \sqrt{\frac{b_k V_k}{\delta B}} Q^{-1}(\varepsilon_k)\right], \quad (6)$$

where b_k is the allocated bandwidth coefficient of the k^{th} IoT UN, δ is the transmission time interval, and Q^{-1} (.) is the inverse function and $Q(\varepsilon_k) = \frac{1}{\sqrt{2\pi}}\int_{\varepsilon_k}^{\infty} \exp^{\left(\frac{-t^2}{2}\right)} dt$, where ε_k is decoding error probability. Here, the channel dispersion is given by $V_k = 1 - [1+\gamma_k]^{-2}$. The uplink transmission latency is given as $T_k^{\text{co}} = \frac{D_k}{R_k}, \forall R_K \in \{R_k^F, R_k^S\}$, where D_k is data size (bits).

B. DT-assisted active RIS computation model

The k^{th} IoT UN locally executes the β_k proportion of the tasks and offloads the $(1-\beta_k)$ proportion of tasks to the ECS when there is an association between the IoT UN and the ECS. A tuple $J_k = (D_k, C_k, T_k^{max})$ represents the task at the k^{th} IoT UN, where T_k^{max} and C_k are the maximum task latency and the required cycles for computation, respectively.

We model the DT service for local processing as $T_k^{un} = (f_k^{un}, \hat{f}_k^{un})$, where f_k^{un} denotes the estimated processing rate and $\hat{f}_k^{un}$ represents the error in the processing rate estimation of the k^{th} IoT UN node.

The processing rate of node P is expressed as

$$T_k^P = \widetilde{T}_k^P + T_{ek}^P = \zeta_k C_k / f_{ek}^P, \quad (7)$$

where $P \in \{un, ecs\}$, $\zeta_k = \beta_k$ for $P = un$ and $\zeta_k = 1 - \beta_k$ for $P = ecs$. Here, $\widetilde{T}_k^P = \zeta_k C_k / f_k^P$ is estimated processing latency, $T_{ek}^P = \frac{\zeta_k C_k \hat{f}_k^P}{[f_k^P (f_{ek}^P)]}$ is the latency error, and $f_{ek}^P = f_k^P - \hat{f}_k^P$ is the processing rate error.

C. Energy and Latency computation

The expressions for energy consumption of the k^{th} IoT UN are given as follows:

$$E_k^{cp} = \beta_k \theta C_k \left(f_k^{un} - \hat{f}_k^{un}\right)^2 /2, \quad (8)$$

$$E_k^{cm} = (1-\beta_k)\, p_k D_k / R_k^F, \quad (9)$$

$$E_k^{tot} = (1-\mu_k)\left[E_k^{cp} + E_k^{cm}\right], \quad (10)$$

where E_k^{cp}, E_k^{cm}, E_k^{tot} and θ are the energy for computation, the energy for communication, the total energy consumption and its power parameter, respectively. Here, $\boldsymbol{\mu} \triangleq \{\mu_k\}\,|\mu_k \in \{0,1\}, \forall k$, represents the IoT UN affiliation with the ECS. When $\mu_k = 1$, there exists a connection between IoT UN and ECS and thereby, the task J_k is cached at the ECS, which is offloaded from the IoT UN, and when $\mu_k = 0$, there exists no connection between IoT UN and ECS [19].

The cumulative latency considering MEC is given as follows:

$$T_k^{tot} = \frac{\mu_k C_k}{f_{ek}^{ecs}} + (1-\mu_k) \times [T_k^{un} + T_k^{\text{co}} + T_k^{ecs}]. \quad (11)$$

D. Problem formulation for Fully-connected active RIS

Our aim in this subsection is to solve the total latency minimization problem by dealing with the optimization of the active RIS phase shift matrix, allocated bandwidth, offloading proportions, MEC policies, transmit power, estimated processing rates at IoT UN and ECS, the energy consumption of IoT UN and MEC capacity at ECS. Thus, the optimization problem for the proposed network in fully-connected architecture can be formulated as follows:

$$\min_{\substack{\mathbf{w}_k,\beta_k,\mu_k,b_k \\ p_k,\mathbf{\Phi},f_k^{\text{un}},f_k^{\text{ecs}}}} \sum_{k=1}^{K} T_k^{\text{tot}}\left(\beta_k,\mu_k,p_k,b_k,\mathbf{\Phi},f_k^{\text{un}},f_k^{\text{ecs}}\right), \quad (12a)$$

$$\text{s.t.} \quad T_k^{\text{tot}}\left(\beta_k,\mu_k,p_k,b_k,\mathbf{\Phi},f_k^{\text{un}},f_k^{\text{ecs}}\right) \le T_k^{\max}, \forall k, \quad (12b)$$

$$\left\|\mathbf{h}_{r,k}\mathbf{\Phi}\right\|^2 p_k + \sigma_0^2 \left\|\mathbf{\Phi}\right\|_F^2 \le P_B/K, \forall k, \quad (12c)$$

$$\sum_{k=1}^{K}\left[\mu_k f_k^{\text{ecs}} + (1-\mu_k)(1-\beta_k) f_k^{\text{ecs}}\right] \le F_{\max}^{\text{ecs}}, \quad (12d)$$

$$E_k^{\text{tot}}\left(\mu_k,\beta_k,f_k^{\text{un}},p_k,b_k,\mathbf{\Phi}\right) \le E_k^{\max}, \forall k, \quad (12e)$$

$$R_k^F\left(p_k,b_k,\mathbf{\Phi}\right) \ge R_{\min}, \forall k, \quad (12f)$$

$$\sum_{k=1}^{K} b_k \le 1, \forall k, \quad (12g)$$

$$\sum_{k=1}^{K} \mu_k D_k \le S_{\max}^{\text{ecs}}, \quad (12h)$$

$$\alpha_n \le \alpha_{max}, \forall n, \quad (12i)$$

$$\mathbf{p} \in \mathscr{P}, \boldsymbol{\beta} \in \mathscr{B}, \mathbf{f} \in \mathscr{F}, \quad (12j)$$

$$\|\mathbf{w}_k\| = 1, \forall k, \quad (12k)$$

where $\mathscr{P} \triangleq \{p_k, \forall k \mid 0 \le p_k \le P_k^{max}, \forall k\}$, $\mathscr{B} \triangleq \{\beta_k, \forall k \mid 0 \le \beta_k \le 1, \forall k\}$, $\mathscr{F} \triangleq \{\mathbf{f} = \{f_k^{\text{un}}, f_k^{\text{ecs}}\}, \forall k | 0 \le f_k^{\text{un}} \le F_{\max}^{\text{un}}, \forall k; 0 \le f_k^{\text{ecs}} \le F_{\max}^{\text{ecs}} \forall k\}$ The constraints are associated with the uplink power, the offloading decisions, and the processing rates. Constraint (12b) indicates the upper limit of the latency. Constraints on active RIS, maximum computing, and energy of the IoT UN are given by (12c), (12d), and (12e), respectively. Constraints (12f),(12g), (12h), and (12i) define the minimum transmission rate, bandwidth allocation of IoT UN, the

caching capacity of the ECS, and the maximum amplitude coefficient, respectively. Note that the $\sum_k E_k^{\max} = E_{\text{ECS}}^{\max}$, where $E_{\text{ECS}}^{\max}$ is the maximum energy consumption at the ECS.

E. Problem formulation for Sub-connected active RIS

In an effort to address the substantial power consumption of fully-connected active RIS, we introduce the sub-connected architecture as a proposed solution [14]. The number of power amplifiers required are denoted as $L = \frac{N}{T}$, where each power amplifier serves T RIS elements in the sub-connected active RIS architecture. The precoding matrix at the sub-connected active RIS can be described as $\boldsymbol{\Psi} = \text{diag}(\boldsymbol{\psi}) = \text{diag}(\boldsymbol{\Phi\Gamma a})$ with $\boldsymbol{\Psi} \in \mathbb{C}^{N\times N}$, where $\boldsymbol{\Gamma} \in \mathbb{C}^{N\times L}$ and $\boldsymbol{a} \in \mathbb{R}^{L\times 1}$ denote the connection indicator matrix for power amplifiers and phase shift circuits, and the amplification factor vector, respectively.

The power consumed by all system components can be written as

$$P_{tot} = \sum_{k=1}^{K} p_k + \xi\left(\sum_{k=1}^{K} \left\|\mathbf{w}_k^H \mathbf{H}_{r,b} \boldsymbol{\Psi}\right\|^2 + \|\boldsymbol{\Psi}\|^2 \sigma_0^2\right) + KP_{\text{U}} + P_{\text{BS}} + NP_{\text{PS}} + LP_{\text{PA}}, \tag{13}$$

where P_{PS} and P_{PA} constitute the hardware static power of active RIS and relate to the phase shift circuit and power amplifier, respectively. P_{U} is the power loss at each IoT UN, ξ is the energy conversion efficiency, and P_{BS} denotes the power loss at the BS.

The latency minimization problem for the proposed network in sub-connected architecture can be formulated as follows

$$\min_{\substack{\mathbf{w}_k,\beta_k,\mathbf{a},\mu_k,b_k \\ p_k,\boldsymbol{\Phi},f_k^{\text{un}},f_k^{\text{ecs}}}} \sum_{k=1}^{K} T_k^{\text{tot}}\left(\beta_k, \mu_k, p_k, b_k, \boldsymbol{\Phi}, f_k^{\text{un}}, f_k^{\text{ecs}}\right), \tag{14a}$$

$$\text{s.t.} \quad \xi\left(\sum_{k=1}^{K} \left\|\mathbf{w}_k^H \mathbf{H}_{r,b} \boldsymbol{\Psi}\right\|^2 + \|\boldsymbol{\Psi}\|^2 \sigma_0^2\right) + NP_{\text{PS}} + LP_{\text{PA}} \le P_{RIS}^{max}, \tag{14b}$$

$$|\phi_n| = 1, \forall n \in [N], \tag{14c}$$

$$a_l \ge 0, \forall l \in [L], \tag{14d}$$

$$(12b), (12d), (12e), (12f), (12g), (12h), (12j), (12k). \tag{14e}$$

where (14b) denotes the constraint of the maximum total power consumed at the active RIS (P_{RIS}^{max}). The feasibile sets of the phase-shift matrix and the amplification factor vector is defined by the (14c) and (14d), respectively.

III. Proposed Solution for Fully-Connected active RIS

The problem in (12) is inherently complex due to the tight coupling of continuous and integer variables, resulting in a challenging non-convex objective function. Therefore, we introduce an alternative optimization algorithm to resolve the issue. This method involves sequentially optimizing individual variables while maintaining the remaining variables constant. We have segregated the original problem into five sub-problems: beamforming design, caching policy optimization, offloading policy optimization, joint computation and communication optimization, and active RIS phase shift optimization.

A. Beamforming Design

In this subsection, we determine the beamforming design as

$$\min_{\substack{\mathbf{w}_k|\boldsymbol{\mu}^{(j)},\boldsymbol{\beta}^{(j)},\mathbf{f}^{(j)},\mathbf{b}^{(j)}, \\ \mathbf{p}^{(j)},\boldsymbol{\Phi}^{(j)}}} \sum_{k=1}^{K} T_k^{\text{tot}}\left(\mu_k\right), \tag{15a}$$

$$\text{s.t.} \quad (12b), (12c), (12f), (12k), \tag{15b}$$

where $f \in \{f_k^{un}, f_k^{ecs}\}$, $\boldsymbol{\mu} \in \mu_k$, $\mathbf{b} \in b_k$, $\mathbf{p} \in p_k, \forall k$ and $\boldsymbol{\beta} \in \beta_k$. To maximize the SNR of each IoT UN, the linear minimum mean-squared error (MMSE) receiver[3] is widely recognized as the most efficient receive beamformer [37]. Hence, we adopt the MMSE-based equalizer for IoT UN which is mathematically expressed as

$$\mathbf{w}_k^* = \frac{\left(\mathbf{h}_k \mathbf{h}_k^H p_k + \mathbf{H}_{r,b} \boldsymbol{\Phi}\boldsymbol{\Phi}^H \mathbf{H}_{r,b}^H \sigma_0^2 + \sigma_b^2 I_M\right)^{-1} \mathbf{h}_k \sqrt{p_k}}{\left\|\left(\mathbf{h}_k \mathbf{h}_k^H p_k + \mathbf{H}_{r,b} \boldsymbol{\Phi}\boldsymbol{\Phi}^H \mathbf{H}_{r,b}^H \sigma_0^2 + \sigma_b^2 I_M\right)^{-1} \mathbf{h}_k \sqrt{p_k}\right\|}. \tag{16}$$

B. Caching Policy Optimization

In this subsection, we aim at determining the subsequent iteration point $\boldsymbol{\mu}^{(j+1)}, \forall \boldsymbol{\mu} \in \{\mu_1, \mu_2 \dots \mu_k\}$ under all the other constraints are fixed. Thus, the subproblem is expressed as

$$\min_{\substack{\mu_k \in \{0,1\}|,\mathbf{w}_k^{(j+1)},\boldsymbol{\beta}^{(j)},\mathbf{f}^{(j)}, \\ \mathbf{b}^{(j)},\mathbf{p}^{(j)},\boldsymbol{\Phi}^{(j)}}} \sum_{k=1}^{K} T_k^{\text{tot}}\left(\mu_k\right), \tag{17a}$$

$$\text{s.t.} \quad (12b), (12d), (12e), (12h). \tag{17b}$$

Since μ_k is an integer variable, this problem exhibits non-convex characteristics. To address this, we introduce the variable $t_k^s = T_k^{un} + T_k^{co} + T_k^{ecs}, \forall k$, and arrange the values of t_k^s in descending order. Subsequently, we prioritize caching the tasks with higher t_k^s until the constraint (12h) is violated in relation to other constraints, allowing us to figure out the optimal values of $\boldsymbol{\mu}$ at the i^{th} iteration, and this may require a few constraints checks ($< K$) to execute optimization of μ [4], [36].

C. Offloading Policy Optimization

With the objective of identifying the next iteration point $\boldsymbol{\beta}^{(j+1)}$, the subproblem can be formulated as follows:

$$\min_{\substack{\beta_k \in [0,1]|\mathbf{w}_k^{(j+1)},\boldsymbol{\mu}^{(j+1)}, \\ \mathbf{f}^{(j)},\mathbf{b}^{(j)},\mathbf{p}^{(j)},\boldsymbol{\Phi}^{(j)}}} \sum_{k=1}^{K} T_k^{\text{tot}}\left(\beta_k\right), \tag{18a}$$

$$\text{s.t.} \quad (12b), (12d), (12e), (12j). \tag{18b}$$

Since this problem consists of linear constraints, it possesses convex characteristics and can be effectively solved using CVX.

[3]The MMSE-based equalizer for the k^{th} IoT UN incorporates the considerations for the increased interference caused by the active RIS noise amplification [39], [40].

D. Joint Communication and Computation Optimization

In this subsection, our objective is to determine the subsequent iteration points $\mathbf{f}^{(j+1)}, \mathbf{b}^{(j+1)}, \mathbf{p}^{(j+1)}$ by fixing the values of $\boldsymbol{\mu}^{(j+1)}$, $\boldsymbol{\beta}^{(j+1)}$ and $\boldsymbol{\Phi}^{(j)}$. Utilizing the approximation $V_k \approx 1$ under the condition of high received SNR [4], R_k^F can be defined as

$$R_k^F \approx \frac{B}{\ln 2}\left[b_k \ln\left(1+\gamma_k\right) - \sqrt{\frac{b_k}{\psi B}}Q^{-1}(\varepsilon_k)\right] \triangleq \frac{B}{\ln 2}\left[\mathbb{G}_k - \mathbb{B}_k\right], \tag{19}$$

where $\mathbb{G}_k = b_k \ln(1+\gamma)$ and $\mathbb{B}_k = \sqrt{b_k}\frac{Q^{-1}(\varepsilon_k)}{\sqrt{\psi B}}$. After considering Taylor's approximation, we reformulate $\mathbb{G}_k$ as

$$\mathbb{G}_k^{(j)} = z\ln\Big(1+\frac{\hat{x}}{\hat{y}}\Big) + x\Big(\frac{\hat{z}}{\hat{x}+\hat{y}}\Big) - y\Big(\frac{\hat{x}\hat{z}}{\hat{y}(\hat{x}+\hat{y})}\Big). \tag{20}$$

Moreover, by using the inequality $\sqrt{z} \leq \frac{\sqrt{\hat{z}}}{2} + \frac{z}{2\sqrt{\hat{z}}}$, we can approximate $\mathbb{B}_k^{(j)}$ as $\mathbb{B}_k \leq \frac{Q^{-1}(\varepsilon_k)}{\sqrt{\psi B}}(\frac{\sqrt{\hat{z}}}{2} + \frac{z}{2\sqrt{\hat{z}}}) \triangleq \mathbb{B}_k^{(j)}$, where $z = b_k$, $x = |\mathbf{w}_k^H\hat{\mathbf{h}}_k|^2 p_k$, $y = ||\mathbf{w}_k^H\mathbf{H}_{r,b}\boldsymbol{\Phi}||^2\sigma_0^2 + ||\mathbf{w}_k^H||^2 B b_k N_0$, $\hat{z} = \hat{b}_k$, $\hat{x} = |\mathbf{w}_k^H\hat{\mathbf{h}}_k|^2\hat{p}_k$, and $\hat{y} = ||\mathbf{w}_k^H\mathbf{H}_{r,b}\boldsymbol{\Phi}||^2\sigma_0^2 + ||\mathbf{w}_k^H||^2 B\hat{b}_k N_0$. Thus, the rate expression is given as $R_k^F \geq R_k^{(j)} \triangleq \frac{B}{\ln 2}\left[\mathbb{G}_k^{(j)} - \mathbb{B}_k^{(j)}\right]$. Now, the approximate expression of constraint (12g) is $R_k^{(j)} \geq R_{min}, \forall k$. To deal with the non-convex constraint (12f), we introduce a new variable $\boldsymbol{\tau_k} \triangleq \{\tau_k\}, \forall k$ such that $1/R_k^F \leq \tau_k, \forall k$ and thus, we can reformulated the constraint (12e) as

$$1/R_k^{(j)} \leq \tau_k, \tag{21}$$

$$(1-\mu_k^{(j+1)})\left[\frac{\theta}{2}\beta_k^{(j+1)}C_k(f_{ek}^{un})^2, + \left(1-\beta_k^{(j+1)}\right)p_k\tau_k\right] \leq E_k^{max}, \forall k. \tag{22}$$

Now that constraint (21) is convex. However, it is noted that (22) is still non-convex. Hence, we approximate the constraint using the inequality given in [4] as follows

$$\left(1-\mu_k^{(j+1)}\right)\left[\frac{\theta\beta_k^{(j+1)}C_k(f_{ek}^{un})^2}{2} + \frac{\left(1-\mu_k^{(j+1)}\right)}{2}\left(\frac{\tau_k^{(j)}}{p_k^{(j)}}p_k^2 + \frac{p_k^{(j)}}{\tau_k^{(j)}}\tau_k^2\right)\right] \leq E_k^{\max}, \forall k. \tag{23}$$

Then, the non-convex objective function (14a) can be approximately represented as follows

$$T_k^{tot} \leq (1-\mu_k^{(j+1)})\left[\frac{\beta_k^{(j+1)}C_k}{f_{ek}^{un}} + D_k\tau_k + \frac{(1-\mu_k^{(j+1)})C_k}{f_{ek}^{ecs}}\right] + \frac{\mu_k^{(j+1)}C_k}{f_{ek}^{ecs}} \triangleq \acute{T}_k^{tot}. \tag{24a}$$

Finally, we formulated the joint communication and computation optimization subproblem as

$$\min_{\mathbf{b},\mathbf{p},\mathbf{f}|\mathbf{w}_k^{(j+1)},\boldsymbol{\mu}^{(j+1)},\boldsymbol{\beta}^{(j+1)}} \sum_{k=1}^{K}\acute{T}_k^{tot}, \forall k, \tag{25a}$$

$$\text{s.t.} \quad \acute{T}_k^{tot} \leq T_k^{max}, \tag{25b}$$

$$(12d), (12f), (12g), (12j), (21), (23). \tag{25c}$$

E. Active RIS phase shift optimization

We optimize the phase shift matrix $\boldsymbol{\Phi}$ of active RIS in this subproblem. Hence, we reformulate the objective function by removing the terms that are independent on $\boldsymbol{\Phi}$ as shown below.

$$\min_{\substack{\boldsymbol{\Phi}|\mathbf{w}_k^{(j+1)},\boldsymbol{\mu}^{(j+1)},\boldsymbol{\beta}^{(j+1)},\\ \mathbf{f}^{(j+1)},\mathbf{b}^{(j+1)},\mathbf{p}^{(j+1)}}} \sum_{k=1}^{K}T_k^{co} = \sum_{k=1}^{K}D_k/R_k^F, \tag{26a}$$

$$\text{s.t.} \quad (12b), (12c), (12e), (12f), (12i). \tag{26b}$$

Here, the problem (26a) holds a non-convexity nature because of the objective function and its respective constraints. To solve problem (26a) by satisfying $R_k^F \geq R_k^{(j)}$, we simplify (6) as

$$R_k^F \triangleq A_1\ln\left(1+\gamma_k\right) - A_2 \triangleq A_1\ln\left(1+\lambda_k\right) - A_2, \tag{27}$$

where $A_1 = \frac{B}{\ln 2}b_k$ and $A_2 = \frac{B}{\ln 2}\sqrt{\frac{b_k V_k}{\delta B}}Q^{-1}(\varepsilon_k)$. To handle the non-convex objective function of problem (26a), we introduce new auxiliary variables $\rho_k \triangleq \{\rho_k\}\forall k$ and λ_k. Thus, the problem (26a) is reformulated as follows

$$\min_{\boldsymbol{\Phi}|\mathbf{w}_k^{(j+1)},\boldsymbol{\mu}^{(j+1)},\boldsymbol{\beta}^{(j+1)},\mathbf{f}^{(j+1)},\mathbf{b}^{(j+1)},\mathbf{p}^{(j+1)}} \sum_{k=1}^{K}D_k\rho_k, \tag{28a}$$

$$\text{s.t.} \quad 1/R_k^{(j)} \leq \rho_k, \tag{28b}$$

$$\gamma_k \geq \lambda_k, \forall k, \tag{28c}$$

$$(12c), (12e), (12i). \tag{28d}$$

Furthermore, the optimal RIS phase shift matrix is obtained by solving the problem (28a). Firstly, we define $\mathbf{H}_{r,k} = \text{diag}\left\{|\left[\mathbf{h}_{r,k}\right]_1|^2, ..., |\left[\mathbf{h}_{r,k}\right]_N|^2\right\}$, $\overline{\mathbf{H}}_{r,k} = \text{diag}\left\{\mathbf{H}_{r,k}, 0\right\}$ to solve the constraint (12c). Now, the constraint (12c) can be reformulated as $\sum_{k=1}^{K}\left\|\mathbf{h}_{r,k}\boldsymbol{\Phi}\right\|^2 p_k + \sigma_0^2\left\|\boldsymbol{\Phi}\right\|_F^2 \leq P_B/K = \text{Tr}(\overline{\mathbf{H}}_r\overline{\mathbf{U}})$, where $\overline{H}_r = p_k\overline{H}_{r,k} + \sigma_0^2\overline{I}, \overline{I} = \text{diag}\left\{I_N, 0\right\}, \overline{U} = \overline{\mathbf{u}}\overline{\mathbf{u}}^H$, and $\overline{\mathbf{u}}^H = \left[\phi_1, ..., \phi_N, 1\right]$. To solve constraint (28c), we define $\mathbf{t}_k^H = \mathbf{w}_k^H\mathbf{H}_{r,b}$ and get $\left|\mathbf{w}_k^H(\mathbf{H}_{r,b}\boldsymbol{\Phi}\mathbf{h}_{r,k})\right|^2 = \text{Tr}(\overline{\mathbf{J}}_k\overline{\mathbf{U}})$, where $\overline{\mathbf{J}}_k = j_k j_k^H, j_k = \text{diag}\left\{\mathbf{o}_k^H\right\}\mathbf{h}_{r,k}$. Defining $\mathbf{O}_k = \text{diag}\left\{|\left[\mathbf{o}_k\right]_1|^2, ..., |\left[\mathbf{o}_k\right]_N|^2\right\}, \overline{\mathbf{O}}_k = \text{diag}\left\{\mathbf{O}_k, 0\right\}$, we have $\sigma_0^2\left\|\mathbf{w}_k^H\mathbf{G}\boldsymbol{\Phi}\right\|^2 + \sigma_b^2\left\|\mathbf{w}_k\right\| = \sigma_0^2\text{Tr}(\overline{\mathbf{O}}_k\overline{U}) + \sigma_b^2\left\|\mathbf{w}_k\right\|$. The optimization of problem (28a) can be transformed as

$$\min_{\boldsymbol{\Phi}} \quad \sum_{k=1}^{K}D_k\rho_k, \tag{29a}$$

$$\text{s.t.} \quad \text{Tr}(\overline{\mathbf{H}}_r\overline{\mathbf{U}}) \leq P_B/K, \tag{29b}$$

$$\text{Tr}(\overline{\mathbf{J}}_1\overline{\mathbf{U}}) \geq ... \geq \text{Tr}(\overline{\mathbf{J}}_K\overline{\mathbf{U}}), \tag{29c}$$

$$\frac{(\overline{\mathbf{J}}_k\overline{\mathbf{U}})p_k}{\sigma_0^2(\overline{\mathbf{T}}_K\overline{\mathbf{U}}) + \sigma_b^2\left\|\mathbf{w}_k\right\|} \geq \lambda_k, \tag{29d}$$

$$\left[\overline{\mathbf{U}}\right]_{n,n} \leq \alpha_{max}^2, \forall n, \tag{29e}$$

$$\overline{\mathbf{U}} \succeq 0, \tag{29f}$$

$$\left[\overline{\mathbf{U}}\right]_{N+1,N+1} = 1, \tag{29g}$$

$$\text{rank}(\overline{\mathbf{U}}) = 1. \tag{29h}$$

The problem (29a) is certainly non-convex due to constraint (29h). Hence, the problem can be transformed into a typical convex semi-definite programming problem utilizing the semi-definite relaxation (SDR) approach. In this way, the constraint (29h) is relaxed, and then it can be solved by using CVX. We adopt the eigenvalue decomposition to obtain a rank-one solution [37]. Suboptimal rank-one solution $\overline{u}^{**}$ can be derived by $\overline{u}^{**} = \sqrt{\mathcal{Y}_i}\mathbf{G}_i$, where $\mathbf{G}_i$ is the eigenvector related with the largest eigenvalue and $\mathcal{Y}_i$ is the largest eigenvalue of $\overline{\mathbf{U}}^*$.

IV. Proposed Solution for Sub-Connected active RIS

The procedure followed in the proposed solution for fully-connected active RIS remains the same for sub-connected active RIS[4], with the key distinction being that the phase shift optimization in the fully-connected architecture is substituted with a joint phase shift and amplification factor vector optimization problem in the sub-connected architecture.

A. Joint phase shift and amplification factor vector optimization

In this subsection, we optimize the phase shift matrix and amplification factor vector of sub-connected active RIS to solve this problem by considering the rate expression as follows

$$\max_{\boldsymbol{\Phi},\boldsymbol{a}} \; R_k^S, \tag{30a}$$

$$\text{s.t.(14b), (14c), (14d).} \tag{30b}$$

Since the (30) is still nonconvex, we introduce $\boldsymbol{\kappa} \in \mathbb{C}^{K\times 1}$ and $\boldsymbol{\nu} \in \mathbb{C}^{K\times 1}$, which are referred to as auxiliary variables. Now, we can reformulate the problem (30) in terms of auxiliary variables as follows:

$$\max_{\boldsymbol{\Phi},\boldsymbol{a},\boldsymbol{\kappa},\boldsymbol{\nu}} \; g(\boldsymbol{\Phi},\boldsymbol{a},\boldsymbol{\kappa},\boldsymbol{\nu}), \tag{31a}$$

$$\text{s.t. (14b), (14c), (14d),} \tag{31b}$$

where

$$\begin{aligned} &g(\boldsymbol{\Phi},\boldsymbol{a},\boldsymbol{\kappa},\boldsymbol{\nu}) \\ &= \sum_{k=1}^{K}\Big[\ln(1+\kappa_k) - \kappa_k + 2\sqrt{1+\kappa_k}\,\mathrm{Re}\left\{\nu_k^*\mathbf{w}_k^{H*}\hat{\mathbf{h}}_k\right\} \\ &\quad - |\nu_k|^2\left(\left\|\mathbf{w}_k^{H*}\mathbf{H}_{r,b}\boldsymbol{\Psi}\right\|^2\sigma_0^2 + \left\|\mathbf{w}_k^{H*}\right\|^2 Bb_kN_0\right)\Big]. \end{aligned} \tag{32}$$

Here, we use AO to alternately optimize the variables $\boldsymbol{\Theta}, \boldsymbol{a}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\nu}$, with all the remaining variables constant. By setting $\partial g/\partial\kappa_k$ and $\partial g/\partial\nu_k$ to 0, $\forall k \in [K]$, we calculate the optimal values of the optimization variables of problem (31a) as given below

$$\kappa_k^{\mathrm{opt}} = \frac{\rho_k}{2}\left(\rho_k + \sqrt{\rho_k^2+4}\right), \tag{33}$$

$$\nu_k^{\mathrm{opt}} = \frac{\sqrt{1+\kappa_k}\mathbf{w}_k^{H*}\hat{\mathbf{h}}_k}{\left\|\mathbf{w}_k^{H*}\mathbf{H}_{r,b}\boldsymbol{\Psi}\right\|^2\sigma_0^2 + \left\|\mathbf{w}_k^{H*}\right\|^2 Bb_kN_0}, \tag{34}$$

[4]Note that we have to replace R_k^F with R_k^S to solve the sub-problems.

where $\rho_k = \mathrm{Re}\left\{\nu_k^*\mathbf{w}_k^{H*}\hat{\mathbf{h}}_k\right\}$.

For the optimal RIS precoding, for ease of notation, let us denote $\boldsymbol{\varpi}_k = \mathbf{w}_k^{H*}\boldsymbol{H}_{r,b}$. Then, $\mathbf{w}_k^{H*}\hat{\mathbf{h}}_k$ can be rewritten as

$$\mathbf{w}_k^{H*}\hat{\mathbf{h}}_k = \mathbf{h}_{r,k}^H\,\mathrm{diag}\left(\boldsymbol{\varpi}_k\right)\boldsymbol{\psi}. \tag{35}$$

The following represents the reformulated active RIS precoding optimization problem by substituting (35).

$$\max_{\boldsymbol{\Theta},\boldsymbol{a}} \quad \mathrm{Re}\left\{\boldsymbol{\psi}^H\boldsymbol{v}\right\} - \boldsymbol{\psi}^H\boldsymbol{Q}\boldsymbol{\psi}, \tag{36a}$$

$$\text{s.t.} \quad \boldsymbol{\psi}^H\boldsymbol{R}\boldsymbol{\psi} \le \widetilde{P}_{\mathrm{RIS}}^{\max}, \tag{36b}$$

$$(14c), (14d), \tag{36c}$$

where

$$\boldsymbol{Q} = \sum_{k=1}^{K}|\nu_k|^2\sigma_z^2\,\mathrm{diag}\left(\boldsymbol{\varpi}_k \odot \boldsymbol{\varpi}_k^*\right), \tag{37}$$

$$\boldsymbol{v} = \sum_{k=1}^{K}\mathrm{diag}\left(\boldsymbol{h}_{r,k}\right)\left(2\sqrt{1+\kappa_k^H}\nu_k^*\boldsymbol{\varpi}_k\right), \tag{38}$$

$$\boldsymbol{R} = \sum_{k=1}^{K}\mathrm{diag}\left(\boldsymbol{\varpi}_k \odot \boldsymbol{\varpi}_k^*\right) + \sigma_0^2\boldsymbol{I}_N. \tag{39}$$

Now, the problem (36c) can be effectively solved by alternating direction method of multipliers [14] as it is a quadratic constraint quadratic programming problem. By considering the constraints (14c) and (14d), amplification factor vector $\boldsymbol{a}^{\mathrm{opt}}$ and the associated phase-shift matrix Θ^{opt} are given by

$$\boldsymbol{a}^{\mathrm{opt}} = \boldsymbol{\Gamma}^{\dagger}\,\mathrm{diag}\left(\exp\left(-j\arg\left(\boldsymbol{\psi}^{\mathrm{opt}}\right)\right)\right)\psi^{\mathrm{opt}}, \tag{40}$$

$$\Theta^{\mathrm{opt}} = \mathrm{diag}\left(\exp\left(j\arg\left(\boldsymbol{\psi}^{\mathrm{opt}}\right)\right)\right). \tag{41}$$

The solutions derived from solving the subproblems, namely (15), (17), (18), and (25), for the fully-connected architecture remain valid when applied to the sub-connected architecture. This is because, in the sub-connected architecture, the additional optimization variable $\mathbf{a}$ is treated as a constant while solving these subproblems.

B. Unified Alternating Optimization (AO) Algorithm

To provide more clarity, the overall procedure to solve the problems (12) and (14) is provided in **Algorithm 1**. The proposed algorithm addresses two distinct cases: fully-connected RIS and sub-connected RIS. The algorithm begins by initializing key variables and setting convergence parameters. In each iteration, the algorithm alternates between updating the fully-connected RIS parameters (phase shifts) upon selection of case 1 and the sub-connected RIS parameters (phase shifts and amplitude coefficients) upon selection of case 2. For the fully-connected case, the algorithm optimizes the RIS phase shifts using an AO approach. Conversely, for the sub-connected RIS, the algorithm optimizes both the phase shifts and amplitude coefficients. The optimization process involves solving convex subproblems for the transmit beamforming vector, RIS phase shifts, user association, and RIS configuration. The algorithm iteratively refines these variables until convergence or a maximum iteration threshold is reached. This algorithm provides a comprehensive approach to optimizing either fully-connected or sub-connected RIS configuration.

Algorithm 1 Unified A0-based Algorithm for Fully-Connected and Sub-Connected RIS

Initialize $j = 0$ and $J^{max} = 50$;
Set initial feasible points $\mathbf{w}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\mathbf{b}^{(0)}$, $\mathbf{p}^{(0)}$, $\mathbf{f}^{(0)}$, $\mathbf{a}^{(0)}$, and $\boldsymbol{\Phi}^{(0)}$;
Set the tolerance $\varepsilon_k = 10^{-3}$;
repeat
    Case 1:Fully-Connected RIS Optimization:
    Solve (15) for given $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $\mathbf{w}^{\star}$ and then update $\mathbf{w}^{(j+1)} = \mathbf{w}^{\star}$;
    Solve (17) for given $\mathbf{w}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $\boldsymbol{\mu}^{\star}$ and then update $\boldsymbol{\mu}^{(j+1)} = \boldsymbol{\mu}^{\star}$;
    Solve (18) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $\boldsymbol{\beta}^{\star}$ and then update $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{\star}$;
    Solve (25) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $(\mathbf{f}^{\star}, \mathbf{b}^{\star}, \mathbf{p}^{\star})$ and then update $(\mathbf{f}^{(j+1)}, \mathbf{b}^{(j+1)}, \mathbf{p}^{(j+1)}) = (\mathbf{f}^{\star}, \mathbf{b}^{\star}, \mathbf{p}^{\star})$;
    Solve (26) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, and $\mathbf{f}^{(j)}$ to find the best solution of $\boldsymbol{\Phi}^{\star}$ then update $\boldsymbol{\Phi}^{(j+1)} = \boldsymbol{\Phi}^{\star}$;
    Case 2: Sub-Connected RIS Optimization:
    Solve (15) for given $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, $\mathbf{a}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $\mathbf{w}^{\star}$ and then update $\mathbf{w}^{(j+1)} = \mathbf{w}^{\star}$;
    Solve (17) for given $\mathbf{w}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, $\mathbf{a}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find the best solution of $\boldsymbol{\mu}^{\star}$ and then update $\boldsymbol{\mu}^{(j+1)} = \boldsymbol{\mu}^{\star}$;
    Solve (18) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\mathbf{f}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, $\mathbf{a}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $\boldsymbol{\beta}^{\star}$ and then update $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{\star}$;
    Solve (25) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$ to find $(\mathbf{f}^{\star}, \mathbf{b}^{\star}, \mathbf{p}^{\star})$ and then update $(\mathbf{f}^{(j+1)}, \mathbf{b}^{(j+1)}, \mathbf{p}^{(j+1)}) = (\mathbf{f}^{\star}, \mathbf{b}^{\star}, \mathbf{p}^{\star})$;
    Solve (30) using $\mathbf{w}^{(j)}$, $\boldsymbol{\mu}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{b}^{(j)}$, $\mathbf{p}^{(j)}$, and $\mathbf{f}^{(j)}$ to find $(\boldsymbol{\Phi}^{\star}, \boldsymbol{a}^{\star})$ and then update $(\boldsymbol{\Phi}^{(j+1)}, \boldsymbol{a}^{(j+1)}) = (\boldsymbol{\Phi}^{\star}, \boldsymbol{a}^{\star})$;
    Increment j value by 1;
until Convergence or $j > J^{max}$.

C. Computational Complexity

TABLE III: Computational Complexity

Variable	Computational Complexity	
	FC	**SC**
$\mathbf{w}_k$	$O\left(K^2\sqrt{4K}\right)$	$O\left(K^2\sqrt{4K}\right)$
$\boldsymbol{\mu}_k$	$O\left(K^2\sqrt{2K+2}\right)$	$O\left(K^2\sqrt{2K+2}\right)$
$\boldsymbol{\beta}_k$	$O\left(K^2\sqrt{3K+1}\right)$	$O\left(K^2\sqrt{3K+1}\right)$
b, p, f	$O\left(16K^2\sqrt{5K+2}\right)$	$O\left(16K^2\sqrt{5K+2}\right)$
$\boldsymbol{\Phi}$	$O\left(N^2\sqrt{4K+N}\right)$	-
$\boldsymbol{\Phi}, \boldsymbol{a}$	-	$O\left((L)^2\sqrt{N+L+1}\right)$

The overall computational complexity for the alternating optimization process in the fully-connected scenario is attributed to the updates of the variables $\mathbf{w}_k$, $\boldsymbol{\beta}_k$, $\boldsymbol{\mu}_k$, $\{\mathbf{b, p, f}\}$, and $\boldsymbol{\Phi}$, as detailed in (15), (17), (18), and (25), respectively. Similarly, in the sub-connected scenario, the computational complexity arises from updating the variables $\mathbf{w}_k$, $\boldsymbol{\beta}_k$, $\boldsymbol{\mu}_k$, $\{\mathbf{b, p, f}\}$ and $\{\boldsymbol{\Phi}, \mathbf{a}\}$, which are explained in (15), (17), (18), (25) and (30), respectively. The computational complexity of optimization subproblem is given by $O\left(V_n^2\sqrt{C_n}\right)$, where V_n and C_n denote the number of scalar variables and the number of linear or quadratic constraints.

Table III categorizes and compares the computational complexities of individual sub problems of the fully-connected and sub-connected algorithms. The overall computational complexity of the fully-connected and sub-connected algorithms can be approximated as $O(K^2\sqrt{4K}+K^2\sqrt{2K+2}+K^2\sqrt{3K+1}+16K^2\sqrt{5K+2}+N^2\sqrt{4K+N})$, and $O(K^2\sqrt{4K} + K^2\sqrt{2K+2} + K^2\sqrt{3K+1} + 16K^2\sqrt{5K+2} + (L)^2\sqrt{N+L+1})$, respectively. In the comparison between fully-connected and sub-connected RIS architectures, the sub-connected approach demonstrates a significant reduction in complexity. This reduction is quantified by a factor of (T^2), highlighting the sub-connected RIS as a more efficient and manageable option in terms of computational complexity.

D. Verification of AO Convergence:

The convergence analysis leverages fixed-point iteration, where each iteration updates a subset of variables, and the mapping $\mathcal{T}$: The minimization process of $f(\mathcal{X})$ is guided by $\mathcal{X}^{(j)} \mapsto \mathcal{X}^{(j+1)}$. If $\mathcal{T}$ is a mapping that has the contraction property, Banach's fixed-point theorem [41] guarantees that $\{\mathcal{X}^{(j)}\}$ converges to a unique fixed point $\mathcal{X}^*$ and $f(\mathcal{X})$ has its stationary solution at this point. Besides that, the convexity of every subproblem also guarantees a monotonic reduction in the objective function which can be given in terms of $f(\mathcal{X}^{(j+1)}) \leq f(\mathcal{X}^{(j)})$. By the compactness of the feasible set and Weierstrass extreme value theorem [42], bound and limit points, $\{\mathcal{X}^{(j)}\}$ are assured. Furthermore, the fact that $f(\mathcal{X})$ is continuously differentiable on its entire domain along with the boundedness of its subgradient shows that any limit point $\mathcal{X}^*$ will be a stationary point, meaning it satisfies first-order necessary conditions for optimality, $\nabla f(\mathcal{X}^*) = \mathbf{0}$. This analysis verifies that the AO algorithm converges to a stationary value under these given conditions, thus eliminating worries about sensitivity to initial values.

V. Numerical Results and Analysis

The proposed system model is investigated under rigorous evaluation through extensive simulations. To ensure the reliability and validity of the simulations, a well-defined set of parameters from [4], [37] is utilized, as listed in **Table IV**.

TABLE IV: Simulation parameters

Parameter	Value	Parameter	Value	Parameter	Value
M, N	2, 4	$F_{\max}^{\text{ecs}}$	30 GHz	$R_{\min}$	0.3 bit/s
K	15	$T_k^{\max}$	10 ms	$E_k^{\max}$	3 mJ
P_B	-5 dBm	B	5 MHz	ε_k	10^{-8}
$d_{b,r}$	150 m	$d_{r,k}$	90 m	P_{RIS}^{max}	9 dBW
$\gamma_{r,k}, \gamma_{b,r}$	[2.6, 2.2]	$Z_{r,k}, Z_{b,r}$	[4, 4]	P_U	10 dBm
P_{BS}	6 dBW	P_{PS}	10 dBm	P_{PA}	10 dBm

Fig. 3 illustrates the convergence of DT-based URLLC systems using active RIS in both fully-connected and sub-connected configurations. This is analyzed within the framework of the proposed AO algorithm, benchmarked against the conventional Heuristic algorithm. From the observations, we observe that the sub-connected RIS exhibits lower latency in both perfect and imperfect CSI conditions when compared to fully-connected active RIS. This trend is rationalized as follows: A fully-connected RIS provides greater control since each element can be individually tuned. However, this

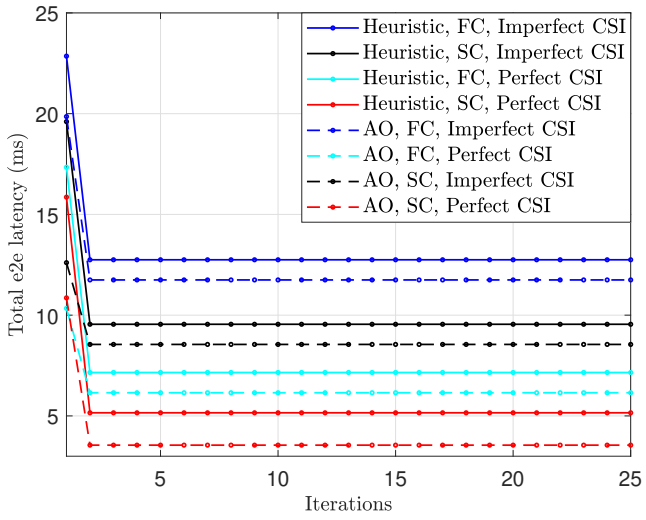

Fig. 3: Convergence of the proposed Algorithm 1 for two distinct cases: fully-connected and sub-connected RIS configurations.

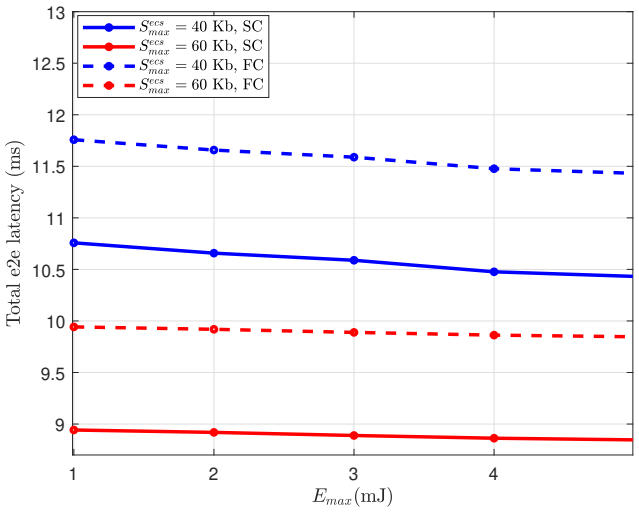

Fig. 4: Maximum energy requirement.

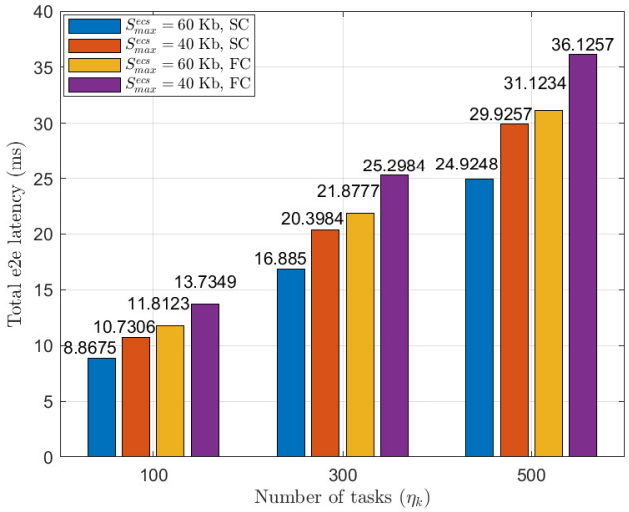

Fig. 5: Impact of task capacity.

necessitates the use of more sophisticated algorithms for configuring these elements, potentially leading to increased processing latency. Conversely, the sub-connected RIS has the potential to reduce processing latency due to fewer active elements requiring control, facilitated by the use of shared power amplifiers for groups of reflection elements. Furthermore, the AO algorithm's performance in these simulations exhibits marked superiority in latency reduction compared to Heuristic approaches. This effectiveness is attributed to the AO algorithm's ability to consistently reach optimal or near-optimal solutions, ensuring a high degree of precision that is particularly beneficial in complex problem-solving scenarios. Additionally, the figure highlights the impact of the initial solution, derived from the SVD-based Heuristic method, on the AO algorithm's convergence. By starting with a solution that approximates the global optimum, the convergence speed and solution quality of the AO algorithm are significantly improved. This aspect is particularly critical in addressing the nonconvex challenges characteristic of DT-based URLLC systems, underscoring the value of a well-conceived initial solution in complex optimization landscapes.

Fig. 4 illustrates the total e2e latency versus E_{max} under different values of $S_{\max}^{\text{ecs}}$. When $S_{\max}^{\text{ecs}}$ increases from 40 Kb to 60 Kb for E_{max} =1 to 5, the total e2e latency decreases. In this context, an increase in edge caching capacity leads to a reduction in latency. Notably, there is a substantial gap between $S_{\max}^{\text{ecs}} = 60$ Kb and $S_{\max}^{\text{ecs}} = 40$ Kb. This pronounced difference serves as clear evidence of the efficacy of the task caching model for time-sensitive metaverse applications in the DT scenario. To illustrate, consider a metaverse application that requires real-time interaction and data exchange. The latency in such an application can be modeled as $T_{\text{metaverse}} = T_{\text{data transmission}} + T_{\text{processing}}$. Our proposed system aims to minimize $T_{\text{data transmission}}$ through efficient beamforming and RIS design, and $T_{\text{processing}}$ through optimized MEC offloading strategies. By achieving a lower $T_{\text{metaverse}}$, our system thus provides a conducive environment for deploying responsive and reliable metaverse-based services. By comparing fully-connected active RIS and sub-connected active RIS, it is observed that the sub-connected active RIS exhibits lower latency. Specifically, the latency of sub-connected active RIS is reduced by 11% compared to fully-connected active RIS under the same scenario (E_{max} = 2 mJ) and ($S_{\max}^{\text{ecs}}$ = 40 Kb). The reason for this trend is explained as follows: The sub-connected active RIS optimizes the RIS phase shift and amplification vector factor, resulting in a stronger signal that reduces data propagation time through the network, consequently lowering latency.

Fig. 5 demonstrates the effect of task capacity on the total e2e latency of sub-connected active RIS as $S_{\max}^{\text{ecs}}$ increases from 40 Kb to 60 Kb. An escalation in task capacity implies more resource-intensive tasks with substantial data transfers, resulting in higher latency. This phenomenon arises because these tasks require more time for completion, causing delays

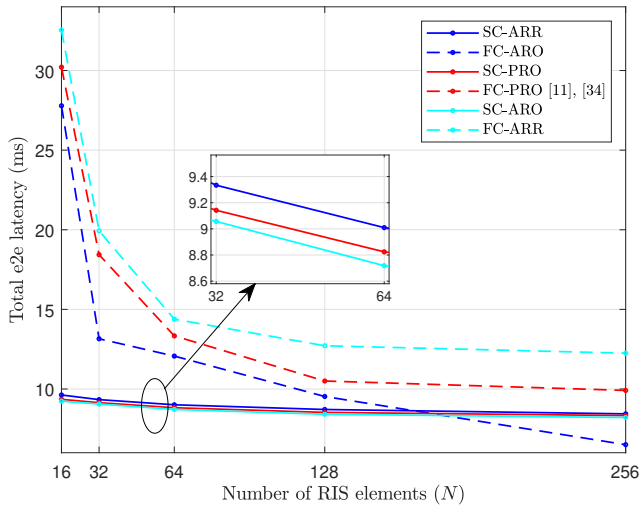

Fig. 6: Effect of N.

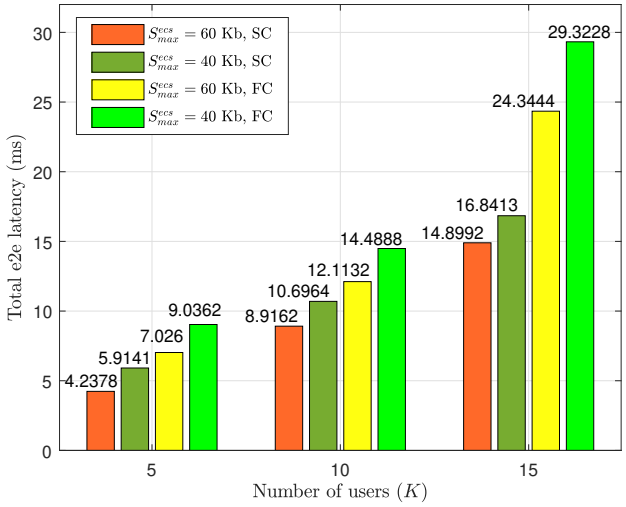

Fig. 8: Effect of K and $S^{\text{ecs}}_{\text{max}}$.

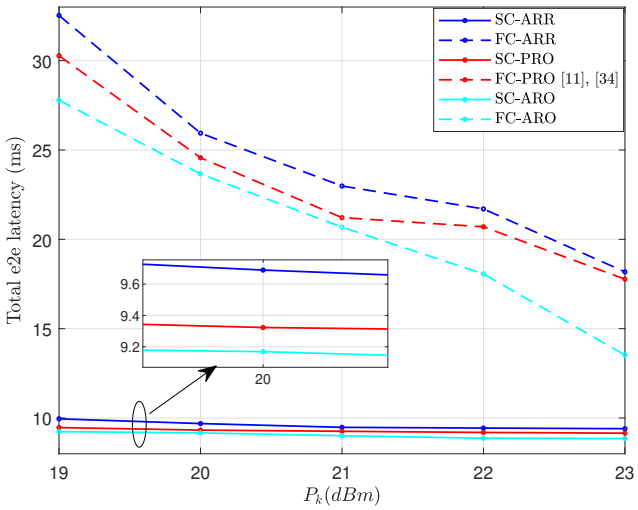

Fig. 7: Effect of P_k.

for IoT UN or systems awaiting a response. Conversely, reduced task capacity signifies less resource demand and lower data requirements, resulting in lower latency, facilitating quicker responses, and enhancing the overall efficiency of the IoT network.

Fig. 6 provides the comparison of the proposed fully-connected (FC) and SC (SC) active RIS schemes with optimal beamforming (ARO) for various values of N against benchmark schemes, such as passive RIS with optimal beamforming (PRO) and active RIS with random beamforming (ARR). The proposed FC and SC configurations using ARO are compared with the existing FC methods using PRO, as detailed in [11] and [34]. The salient findings depicted in the figure can be summarized as follows: By employing higher values of N in both active and passive RIS configurations and adopting a strategy of using fewer active elements for control in the SC setup, several notable advantages are realized. These include an enhancement in channel gain diversity and improved beamforming capabilities. Additionally, this approach leads to a reduction in the computational burden, an improvement in the SNR, and consequently, a significant decrease in latency. Additionally, ARO in the SC configuration achieves a latency reduction of 1.99 times compared to the ARO scheme in the FC configuration. This advantage is attributed to the active RIS intelligent phase adjustment and beamforming optimization capabilities, which effectively mitigate interference. Moreover, when comparing SC PRO to the FC PRO scheme and SC ARR to the FC ARR, significant latency reductions of 2.23 times and 2.38 times, respectively, are achieved. This is because the active RIS with a SC architecture optimizes the amplification factor vector along with the RIS phase shift matrix. A higher amplification factor enhances signal strength, leading to a reduction in data propagation time through the network, consequently lowering latency. Furthermore, enhanced RIS phase adjustment optimizes signal direction, improving the quality of wireless communication links by intelligently changing the radiation propagation environment and thereby reducing latency.

Fig. 7 depicts the impact of P_k on total e2e latency in FC and SC configurations of the proposed network against benchmark schemes, such as PRO and ARR. The proposed FC and SC configurations using ARO are compared with the existing FC methods using PRO, as detailed in [11] and [34]. The observations of Fig. 7 are provided as follows: Increasing the power improves the signal, reduces susceptibility to interference, ensures clearer communications, and reduces delays, resulting in faster data transmission and lower e2e latency. In the FC scenario, SC ARO achieves 2 times lower latency than FC ARO, operating within the same power budget. Similarly, the SC PRO scheme surpasses the FC PRO, and the SC ARR outperforms the FC ARR by achieving latency reductions of 2.19 times and 2.27 times, respectively. The advantage of SC RIS lies in having fewer active elements to control, potentially reducing processing latency.

Fig. 8 reveals the total e2e latency versus K at different

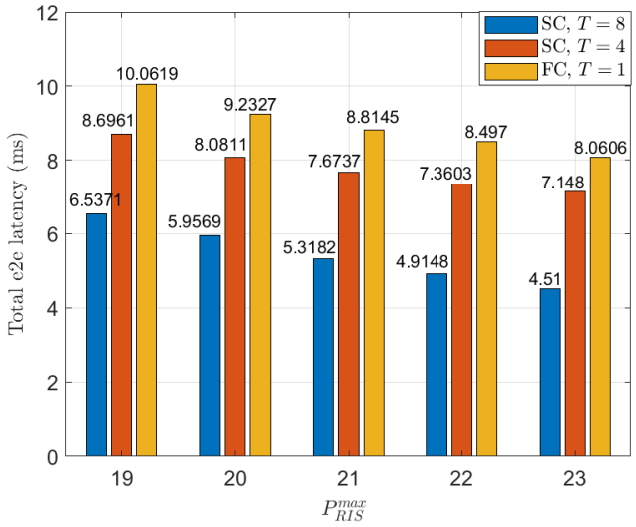

Fig. 9: Maximum power at RIS.

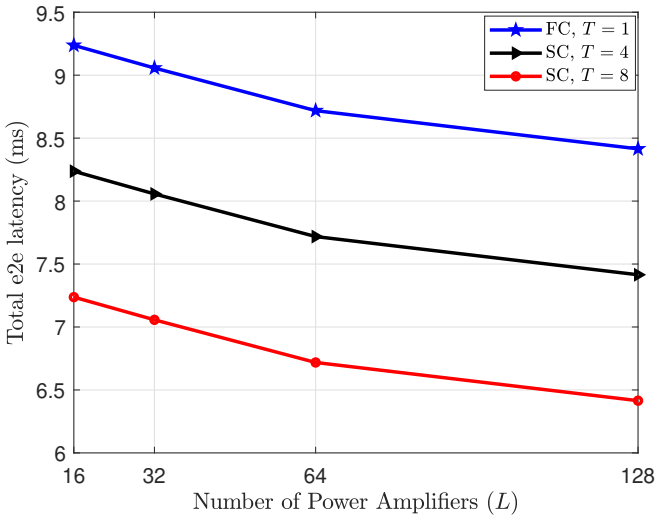

Fig. 10: Number of Power amplifiers.

S_{max}^{ecs} in both fully-connected and sub-connected systems. When the S_{max}^{ecs} increase from 40 Kb to 60 Kb for $K = 5$, $K = 10$, and $K = 15$, it results in a decrease in total e2e latency by 22.24%, 16.39%, and 16.97% for fully-connected and 28%, 16.7%, and 11.5% for sub-connected. We draw two observations from the results. First, the decreasing trend in latency when increasing the S_{max}^{ecs} occurs because higher edge caching capacity allows for more efficient data storage and retrieval in the ECS reducing the need for data transmission from IoT UN to the server. As a result, reduced data transmission requirements lead to lower latency, improving system performance. Second, as K increases, latency increases. This is because, with more IoT devices, there is higher competition for network resources, leading to increased congestion and potential data transmission delays. These findings demonstrate that the sub-connected system exhibits reduced latency, as the sub-connected RIS has the capacity to decrease processing latency by controlling fewer active elements.

Fig. 9 depicts the total end-to-end (e2e) latency versus maximum power of the RIS (P_{RIS}^{max}) under different number of RIS elements served by each power amplifier (T). Specifically, $T = 1$ is employed for the FC scenario, while $T = 4$ and $T = 8$ are used for the sub-connected configuration. The increase in the value of T correlates with a reduction in latency, as observed from the experimental results. Notably, in the sub-connected condition with $T = 8$, latency experiences a decrease of 24% compared to sub-connected condition with $T = 4$. This decrease in latency is primarily due to the more efficient power management facilitated by the sub-connected architecture. In the sub-connected setup, power amplifiers are shared among groups of RIS elements, enhancing energy efficiency while optimizing signal paths and reducing interference. Consequently, signal quality is improved. This study indicates that increasing the value of T in a sub-connected architecture can result in a significant improvement in energy efficiency and signal quality, as reflected in the reduction of latency. These findings strengthen the argument that the sub-connected approach is capable of addressing issues that arise in a FC setup.

Fig. 10 illustrates the total end-to-end (e2e) latency in relation to L for various values of T. The fully-connected scenario is indicated by $T = 1$, while the sub-connected configurations are represented by $T = 4$ and $T = 8$. Experimental data show an inverse relationship between T and latency, highlighted by two significant observations: Firstly, at a fixed value of L, an increase in T leads to a consistent decrease in latency. This trend underlines the superior efficiency of the sub-connected configurations over the fully-connected ones. It aligns with the relationship $L = \frac{N}{T}$, where a higher T leads to a lower effective load per active element, thereby reducing the latency. Secondly, at a fixed value of T, increasing L results in a decrease in latency. This pattern holds true across various T values. A higher L at a fixed T indicates an increased total number of elements N, which contributes to more efficient signal processing and resource allocation, hence reducing latency. For instance, within the fully-connected setup at $T = 1$, a rise in L from 32 to 64 leads to a latency reduction of 3.4%. In the sub-connected scenario at $T = 4$ with $L = 64$, there is a 4.2% decrease in latency compared to $L = 32$. Additionally, in the sub-connected state at $T = 8$ with $L = 64$, the latency reduction is even more significant at 4.8% compared to $L = 32$. These latency improvements are attributed to the efficient power management inherent in the sub-connected architecture. In this setup, power amplifiers are distributed among groups of RIS elements, enhancing energy efficiency, optimizing signal paths, reducing interference, and consequently improving signal quality.

VI. Conclusion

Our research effectively demonstrated the benefits of integrating MEC and sub-connected active RIS-assisted DT communication system in improving task offloading efficiency and reduced latency in edge computing environments for IoT-URLLC services. We proposed a e2e latency optimization

problem for the proposed DT-enabled MEC-URLLC network for both sub-connected and fully-connected architectures, which is solved using AO algorithm and compared against the benchmark scheme i.e., Heuristic approach. We observed that the results illustrate the superior performance of the sub-connected active RIS, surpassing benchmark schemes with significant improvements when compared to the fully-connected active RIS scheme. Specifically, it achieved a performance improvement of 2 times compared to fully-connected ARO, 2.19 times compared to fully-connected PRO, and 2.27 times to fully-connected ARR. Furthermore, we observed that there is a significant reduction in the total e2e latency of the proposed system with an sub-connected active RIS architecture for increased S_{max}^{ecs}, increased number of power amplifiers, increased number of RIS elements and for decreased number of UNs. These significant improvements underscored the inherent advantages of the sub-connected architecture of the proposed system in reducing latency for IoT URLLC services.

REFERENCES

[1] D. Van Huynh, V.-D. Nguyen, S. R. Khosravirad, V. Sharma, O. A. Dobre, H. Shin, and T. Q. Duong, "URLLC edge networks with joint optimal user association, task offloading and resource allocation: A digital twin approach," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7669–7682, Nov. 2022.

[2] M. Gao, R. Shen, L. Shi, W. Qi, J. Li, and Y. Li, "Task partitioning and offloading in DNN-task enabled mobile edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 4, pp. 2435–2445, Sep. 2023.

[3] M. Aloqaily, O. Bouachir, F. Karray, I. A. Ridhawi, and A. E. Saddik, "Integrating digital twin and advanced intelligent technologies to realize the metaverse," *IEEE Consum. Electron. Mag.*, pp. 1–8, Oct. 2022.

[4] D. Van Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1733–1737, Aug. 2022.

[5] Y. Wu, K. Zhang, and Y. Zhang, "Digital twin networks: A survey," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 13 789–13 804, Sep. 2021.

[6] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, Jan. 2023.

[7] G. Szabó, S. Rácz, N. Reider, H. A. Munz, and J. Pető, "Digital twin: Network provisioning of mission critical communication in cyber physical production systems," in *Proc. IEEE Int. Conf. Ind. 4.0 Artif. Intell.Commun. Technol. (IAICT),*, July 2019.

[8] H. Ren, K. Wang, and C. Pan, "Intelligent reflecting surface-aided URLLC in a factory automation scenario," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 707–723, Jan. 2022.

[9] Z. Peng, R. Weng, Z. Zhang, C. Pan, and J. Wang, "Active reconfigurable intelligent surface for mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2482–2486, Dec. 2022.

[10] S. Kurma, K. Singh, M. Katwe, S. Mumtaz, and C.-P. Li, "RIS-empowered MEC for URLLC systems with digital-twin-driven architecture," in *Proc. IEEE Int. Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2023, pp. 1–6.

[11] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Enabling URLLC applications through reconfigurable intelligent surfaces: Challenges and potential," *IEEE Internet Things Mag.*, vol. 5, no. 1, pp. 130–135, Mar. 2022.

[12] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active RIS vs. passive RIS: Which will prevail in 6G?" *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2023.

[13] C. You and R. Zhang, "Wireless communication aided by intelligent reflecting surface: Active or passive?" *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2659–2663, Dec. 2021.

[14] K. Liu, Z. Zhang, L. Dai, S. Xu, and F. Yang, "Active reconfigurable intelligent surface: Fully-connected or sub-connected?" *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 167–171, Jan. 2022.

[15] Q. Zhu, M. Li, R. Liu, Y. Liu, and Q. Liu, "Joint beamforming designs for active reconfigurable intelligent surface: A sub-connected array architecture," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7628–7643, Nov. 2022.

[16] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for secure URLLC in mission-critical IoT scenarios," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5793–5807, Sep. 2020.

[17] W. R. Ghanem, V. Jamali, and R. Schober, "Resource allocation for secure multi-user downlink MISO-URLLC systems," in *in Proc. IEEE Int. Conf. Commun. (ICC)*, June 2020, pp. 1–7.

[18] J. Liu, M. Ahmed, M. A. Mirza, W. U. Khan, D. Xu, J. Li, A. Aziz, and Z. Han, "RL/DRL meets vehicular task offloading using edge and vehicular cloudlet: A survey," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8315–8338, June 2022.

[19] T. Do-Duy, D. Van Huynh, O. A. Dobre, B. Canberk, and T. Q. Duong, "Digital twin-aided intelligent offloading with edge selection in mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 806–810, Apr. 2022.

[20] 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.

[21] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, May 2021.

[22] 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.

[23] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[24] 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. 13 913–13 926, Aug. 2022.

[25] 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.

[26] S. Yang, W. Lyu, D. Wang, and Z. Zhang, "Separate channel estimation with hybrid RIS-aided multi-user communications," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1318–1324, Jan. 2023.

[27] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[28] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–6.

[29] 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.

[30] W. R. Ghanem, V. Jamali, and R. Schober, "Joint beamforming and phase shift optimization for multicell IRS-aided OFDMA-URLLC systems," in *Proc. IEEE Wireless Commun. Netw. Conf*, May 2021, pp. 1–7.

[31] 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.

[32] D. C. Melgarejo, C. Kalalas, A. S. de Sena, P. H. J. Nardelli, and G. Fraidenraich, "Reconfigurable intelligent surface-aided grant-free access for uplink URLLC," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.

[33] X. Gao, W. Yi, Y. Liu, and L. Hanzo, "Multi-objective optimisation of URLLC-based Metaverse services," *IEEE Trans. Commun.*, pp. 1–1, Aug. 2023.

[34] B. Li, Y. Liu, L. Tan, H. Pan, and Y. Zhang, "Digital twin assisted task offloading for aerial edge computing and networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10 863–10 877, Oct. 2022.

[35] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for hybrid 5G services in mobile edge computing systems: Learn from a digital twin," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4692–4707, Oct. 2019.

[36] B. Li, W. Xie, Y. Ye, L. Liu, and Z. Fei, "Flexedge: Digital twin-enabled task offloading for UAV-aided vehicular edge computing," *IEEE Trans. Veh. Technol.*, pp. 1–6, June 2023.

[37] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021.
[38] C. She, C. Yang, and T. Q. Quek, "Radio resource management for ultra-reliable and low-latency communications," *" IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, June 2017.
[39] R. Long, Y.-C. Liang, Y. Pei, and E. G. Larsson, "Active reconfigurable intelligent surface-aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4962–4975, Aug. 2021.
[40] H. Niu, Z. Lin, K. An, J. Wang, G. Zheng, N. Al-Dhahir, and K.-K. Wong, "Active ris assisted rate-splitting multiple access network: Spectral and energy efficiency tradeoff," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1452–1467, May 2023.
[41] C. Nwaigwe and D. N. Benedict, "Generalized banach fixed-point theorem and numerical discretization for nonlinear volterra-fredholm equations," *J. Comput. Appl. Math.*, vol. 425, p. 115019, 2023.
[42] C. R. Alcantud, Jose, "Softarisons: theory and practice," *Neural Computing and Applications*, vol. 33, no. 23, pp. 16 759–16 771, 2021.

Sravani Kurma (Graduate student member, IEEE) received the B.Tech. degree in Electronics and Communication Engineering from the JNTUH college of Engineering, Jagtial, India, in 2017, and Master's degree (Gold Medalist) in Communication System Engineering from Visvesvaraya National Institute of Technology, Nagpur, India, in 2019. She is currently pursuing Ph.D in Institute of Communications Engineering (ICE) in National Sun Yat-sen University, Taiwan. Her current research interests include 5G, 6G, Industrial internet of things (IIoT), wireless energy harvesting (EH), cooperative communications, Reconfigurable intelligent surfaces (RIS), Full-duplex communication, cell-free MIMO, ultra-reliable and low latency communication (URLLC), resource allocation, and machine learning for communication.

Tri Ayu Lestari (Graduate Student Member, IEEE) received the Associate Engineering with Cum Laude for Telecommunication Engineering at State Polytechnic of Padang, Padang, Indonesia on 2019, Bachelor of Applied Science with Cum Laude for Telecommunication Engineering at Electronic Engineering Polytechnic Institute of Surabaya (EEPIS), Surabaya, Indonesia on 2022. She is currently pursuing Master Degree in International Master Program Telecommunication Engineering (IMPTE) at National Sun Yat-Sen University (NSYSU), Kaohsiung, Taiwan. Her current research interests includes reconfigurable intelligent surfaces (RIS) and ultra-reliable and low latency communication (URLLC).

Keshav Singh (**Member, IEEE**) received the M.Sc. degree in Information and Telecommunications Technologies from Athens Information Technology, Greece, in 2009, and the Ph.D. degree in Communication Engineering from National Central University, Taiwan, in 2015. He currently works at the Institute of Communications Engineering, National Sun Yat-sen University (NSYSU), Taiwan as an Assistant Professor. Prior to this, he held the position of Research Associate from 2016 to 2019 at the Institute of Digital Communications, University of Edinburgh, U.K. From 2019 to 2020, he was associated with the University College Dublin, Ireland as a Research Fellow. He had chaired workshops on conferences like IEEE GLOBECOM 2023 and IEEE WCNC, 2024. He also serves as leading guest editor of IEEE Transactions on Green Communications and Networking Special Issue on Design of Green Near-Field Wireless Communication Networks. He leads research in the areas of green communications, resource allocation, transceiver design for full-duplex radio, ultra-reliable low-latency communication, non-orthogonal multiple access, machine learning for wireless communications, integrated sensing and communications, non-terrestrial networks, and large intelligent surface assisted communications.

Anal Paul (**Member, IEEE**) received his B.Tech. degree in Information Technology from the Government College of Engineering and Ceramic Technology, Kolkata, India, in 2008, and a Postgraduate (M.E) degree in Software Engineering (Department of Information Technology) from Jadavpur University, India, in 2010. He received his Ph.D. degree from the Department of Information Technology at the Indian Institute of Engineering Science and Technology, Shibpur, in 2021. He was a postdoctoral researcher at the Department of Information and Communication Engineering, Yeungnam University, South Korea from July 2022 to December 2022. Since January 2023, he has been a postdoctoral researcher at the Institute of Communications Engineering (ICE), National Sun Yat-sen University, Taiwan, focusing on advancing 5G/6G wireless technologies through machine learning applications.

Shahid Mumtaz (**Senior Member, IEEE**) is a Professor at Nottingham Trent University, UK and an IET Fellow, IEEE ComSoc, IAS and ACM Distinguished speaker, recipient of IEEE ComSoC Young Researcher Award (2020), founder and EiC of IET "Journal of Quantum communication," Vice-Chair: Europe/Africa Region- IEEE ComSoc: Green Communications & Computing society and Vice-chair for IEEE standard on P1932.1: Standard for Licensed/Unlicensed Spectrum Interoperability in Wireless Mobile Networks. He is the author of 4 technical books, 12 book chapters, 300+ technical papers (200+ IEEE Journals/transactions, 100+ conference, 2 IEEE best paper award- in the area of mobile communications. Most of his publication is in the field of Wireless Communication. He is serving as Scientific Expert and Evaluator for various Research Funding Agencies. He was awarded an "Alain Bensoussan fellowship" in 2012. He is the recipient of the NSFC Researcher Fund for Young Scientist in 2017 from China.