Exploiting Active STAR-RIS to enable URLLC in Digitally-Twinned Internet-of-Things Networks

Tri Ayu Lestari, Sravani Kurma, *Member, IEEE*, Anal Paul, *Member, IEEE*, Keshav Singh, *Member, IEEE*, Simon L. Cotton, *Senior Member, IEEE*, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—In the context of ultra-reliable low-latency communication (URLLC) in Internet-of-Things (IoT) networks, conventional half-space coverage limits the flexibility of reconfigurable intelligent surface (RIS) deployment. To overcome these constraints, this paper makes use of active simultaneously transmitting and reflecting RIS (STAR-RIS), which is seamlessly integrated into digital twin (DT) and mobile edge computing (MEC) frameworks. Our primary research objective is to achieve full-space coverage by enabling simultaneous transmission and reflection of the signals while improving uplink data transmission from IoT URLLC user nodes (UNs) to the base station (BS) with the assistance of active STAR-RIS, even in the presence of imperfect channel state information (CSI). We formulate the problem of minimizing total end-to-end (e2e) latency, computed using the alternating optimization (AO) algorithm. Subsequently, we have evaluated the performance of the AO algorithm against the stochastic gradient descent (SGD) algorithm, which serves as the benchmark solution. The simulation outcomes delineate a performance evaluation under perfect and imperfect CSI scenarios. The AO algorithm outperforms SGD with latency reductions of 19.7% at $N = 32$ and 20.4% at $N = 64$. Increasing N from 32 to 64 results in a 39.3% latency reduction for AO, surpassing SGD's 38.8%. However, the SGD algorithm consistently exhibits lower computational complexity compared to the AO algorithm. Additionally, the energy splitting mode achieves the system's total e2e latency reductions of 28.4% over the mode switching mode and 11.04% over time switching mode. Furthermore, active STAR-RIS optimal beamforming (ARO) achieves ≈10% latency reduction over the predictive optimal beamforming (PRO), which itself surpasses active STAR-RIS with random beamforming (ARR) by ≈9%. This comparison considers key factors such as the power budget, the number of RIS elements, the caching capacity of the edge computing server (ECS), the number of IoT UNs, the minimum transmission rate, and maximum transmit power at BS of active STAR-RIS.**

***Index Terms*—alternating optimization, digital twin, imperfect CSI, mobile edge computing, simultaneously transmitting and reflecting, ultra-reliable low-latency communication.**

The work of K. Singh was supported by the National Science and Technology Council of Taiwan under Grant NSTC 112-2221-E-110-038-MY3. The work of S. L. Cotton was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the EPSRC Hub on All Spectrum Connectivity (EP/X040569/1 and EP/Y037197/1). The work of T. Q. Duong was supported by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109. (*Corresponding author: Keshav Singh*).

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

S. L. Cotton is with the Centre for Wireless Innovation, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K. (E-mail: simon.cotton@qub.ac.uk).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada and is also with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K. (E-mail: tduong@mun.ca).

I. Introduction

Ensuring connectivity for Internet-of-Things (IoT) is vital for time-critical communication, where the integration of mobile edge computing (MEC) serves to reduce overall network latency [1]. This emergent technology will empower industrial IoT applications that require ultra-reliable and low-latency communications (URLLC), paving the way for a new generation of services and experiences [2], [3]. However, achieving efficient task offloading in edge computing environments is challenged by many factors, such as joint computations, heterogeneous architectures, 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 [4]. DT represents a virtual replica of a physical object, system, or process, enabling simulation, analysis, and optimization. On the other hand, Metaverse refers to an immersive virtual environment that allows physical objects (e.g., IoT user nodes (UNs)) to interact with the virtual/digital world [5]. The integration of DT and MEC with real-time optimization theory enables uninterrupted end-to-end (e2e) Metaverse services, enabling a seamless blend of the physical and virtual worlds [6].

In next-generation wireless services, particularly URLLC, the demand for massive IoT user connectivity and latency-sensitive applications poses significant challenges [2], [7]. To overcome these challenges, the integration of DT with URLLC has gained attention [7], [8]. Integrating the capabilities of DTs with URLLC empowers industries to attain reliable and low-latency communication. This synergy is particularly important for mission-critical applications such as autonomous vehicles and industrial automation. This transformative potential extends across multiple domains, including manufacturing, transportation, healthcare, and more, paving the way for smart industries [7]. Previous works have proposed DT-aided edge network approaches and optimization algorithms to minimize computation latency and improve reliability [7], [8]. Additionally, a few studies [9]–[11] have demonstrated the potential of DT and Metaverse technologies in enhancing task offloading efficiency and latency performance. However, most prior research assumes a direct link to the MEC, which can be impractical due to environmental obstacles that hinder direct connectivity. The use of reconfigurable intelligent surfaces (RISs) can offer alternative transmission paths, helping to sustain wireless links which could otherwise be shadowed

or blocked [10]–[13]. While passive RIS relies exclusively on phase-shift control to manipulate signal reflection, active RIS introduces individual power amplifiers within each RIS element, enabling active signal amplification [14].

Recognizing that spatial confinement may not always be achievable, it can impose substantial limitations on the versatility and efficiency of RIS. Consequently, there has been increasing interest in an emerging concept of simultaneously transmitting and reflecting (STAR-RIS) [15] to achieve comprehensive coverage of the entire space. Active STAR-RIS, in particular, introduces a new degree of freedom in manipulating signal propagation, thereby enhancing the flexibility of network design, faster data transmission, and reduced latency, enabling signals to travel more efficiently through optimized pathways. The synergy between the Metaverse, DT networks, and active STAR-RIS is paving the way for communication technology advancements that offer reliable communication while ensuring minimal delays in mission critical IoT applications, such as industrial IoT automation and IoT-enabled autonomous vehicles.

A. Related works

Earlier research predominantly concentrated on resource allocation for secure URLLC, prioritizing communication aspects while neglecting computation considerations [16], [17]. Nevertheless, with the increasing need for devices in mission-critical applications to perform computation-intensive tasks within strict time constraints, MEC has emerged as a compelling solution to facilitate swift and efficient computation in URLLC systems [9], [18]. At the same time, DT technology provides practical features that enable organizations to meet the stringent requirements of high reliability and low latency, ensuring continuous and dependable operation [7]. In the context of IoT networks, where seamless and reliable connectivity is crucial, the incorporation of DT with URLLC holds particular significance as it streamlines the optimization of configurations to meet URLLC requirements [8]–[11].

In the continuously growing domain of IoT networks amid bustling urban environments, the pursuit of improved connectivity has sparked a surge of innovative research on RIS [19], [20], [20]–[23]. Specifically, recent research demonstrates the effectiveness of active RIS compared to passive RIS [23], [24]. The preceding research has predominantly centered on the functionality of active RIS, which primarily act as either reflective or transmissive metasurfaces, as evidenced by studies [22], [25]. In this context, active RIS requires the served IoT UNs to be positioned on the same side as the RIS, resulting in half-space coverage. This configuration, however, imposes limitations on the flexibility of deploying active RIS, thus prompting the need for further investigation into strategies that enhance deployment flexibility and spatial coverage. 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 [21], [26]–[28]. However, only a limited number of studies have specifically explored the application of RIS in URLLC scenarios [29]–[31]. The authors in [29] 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 (MISO) system, where a group of BSs collaborates to serve URLLC traffic. In [10], [11], [32], the authors proposed a 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, the limited spatial coverage constrains the flexibility of RIS deployment, wherein served IoT UNs must be on the same side as the RIS, necessitating the capability of active STAR-RIS to extend coverage [15]. Secondly, a critical need exists to push the technological frontier forward by leveraging the capabilities of edge intelligence to unveil new potentials in URLLC services within MEC-enabled DT networks [7], [8], [36]. Meeting these requirements is essential to empower industries and applications relying on real-time, mission-critical communication, and to stimulate the growth of IoT-enabled innovations. Our study stands at the intersection of this endeavor, positioned to uncover the trade-offs and advantages of active STAR-RIS architectures in DT-based MEC communication frameworks, laying the groundwork for future research and advancements in this emerging field.

As the IoT technology evolves, the integration of sophisticated wireless technologies supported by industry leaders like Microsoft, Google, Qualcomm, Amazon, and TSMC becomes essential. These technologies, which include edge computing and low-power RF tech tailored for URLLC, enhance infrastructure and enrich IoT functionalities to meet the real-time, high-reliability demands of modern industrial applications. This shift indicates a move towards more capable and efficient IoT systems. Our research aligns with these advancements, showing how the integration of advanced 5G and edge computing is crucial for the next generation of IoT frameworks. We contribute to this field by providing insights into the deployment and benefits of advanced RIS technologies, addressing the complex needs of IoT-enabled industries. The contributions of our work are summarized as:

- Unlike [8]–[11], [37], we investigate an active STAR-RIS-assisted DT-based MEC system to facilitate task offloading and enhance IoT-URLLC services. Our main objective is to minimize total end-to-end (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 IoT UN and edge computing server (ECS), active STAR-RIS phase shift matrices, and allocated bandwidth at each IoT UN.
- We develop an efficient alternating optimization (AO) algorithm to address the proposed non-convex problem by dividing it into manageable subproblems: beamforming design, caching policy optimization, offloading policy

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

Paper	RIS	MEC	URLLC	Algorithm	DT	Performance metric
[7]	✗	✓	✓	AO	✓	Worst-case latency minimization of e2e DT latency
[8]	✗	✓	✓	AO	✓	Total e2e latency minimization
[9]	✗	✓	✗	AO	✓	Latency minimization
[10]	Passive RIS	✓	✓	DRL	✓	Latency minimization
[11]	Passive RIS	✓	✓	DRL, AO	✓	Latency minimization
[14]	Active RIS, Passive RIS	✓	✗	BCD, SCA	✗	Minimize the maximum computational latency
[15]	Active STAR-RIS	✗	✗	AO	✗	Maximized communication sum rate
[16]	✗	✗	✓	AO	✗	Latency minimization
[17]	✗	✗	✓	SCA	✗	Minimization of the total transmit power
[19]	Passive RIS	✗	✗	AO	✓	Rate maximization
[21]	Passive RIS	✗	✗	Dinkelbach's Method	✗	Energy efficiency maximization
[23]	Active RIS, Passive RIS	✗	✗	AO	✗	Sum rate maximization
[24]	Active RIS, Passive RIS	✗	✗	AO	✗	Rate maximization
[25]	Passive STAR-RIS	✗	✗	Search-based algorithm	✗	Rate maximization
[28]	Passive RIS	✗	✗	AO	✗	Minimize the total transmit power
[29]	Passive RIS	✗	✓	SCA	✗	Maximization of the weighted sum throughput
[32]	Passive RIS	✗	✓	SGD, MO-SAC	✗	Latency and the total service cost minimization
[33]	Passive RIS	✗	✓	AO	✗	Average decoding error probability and data rate
[34]	✗	✓	✓	DDN	✓	Energy consumption minimization
[35]	✗	✓	✗	PPO	✓	Energy consumption minimization
Our work	Active STAR-RIS	✓	✓	AO	✓	Total e2e latency minimization

optimization, joint optimization of communication and computation, and active STAR-RIS beamforming optimization. Moreover, we verify the convergence of the AO algorithm. To further enhance the system's performance, we chose stochastic gradient descent (SGD) as our benchmark because it offers a robust and widely recognized method for solving optimization problems, particularly in complex environments like those involving active STAR-RIS. We also provide the complexity analysis of both the AO and SGD algorithms.

- Our simulation results confirm that our proposed active STAR-RIS optimal beamforming (ARO) scheme consistently outperforms benchmark schemes, including passive STAR-RIS with optimal beamforming (PRO) and active STAR-RIS with random beamforming (ARR), FD relay in terms of latency when considering factors such as power budget, number of RIS elements, number of IoT UNs, minimum transmission rate, and STAR-RIS maximum transmit power at the base station and different working modes of active STAR-RIS systems.

The remainder of this paper is organized as follows: Section II explains the proposed system model. Section III presents the proposed solutions for active STAR-RIS, while Section IV provides extensive numerical analysis to demonstrate the effectiveness of the considered network. Finally, in Section V, 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

Our proposed system model employs an active STAR-RIS to enhance a URLLC IoT network, as illustrated in Fig. 1. We integrate a macro base station and small cells to address coverage and interference challenges in dense IoT environments such as industrial settings. This architecture enhances

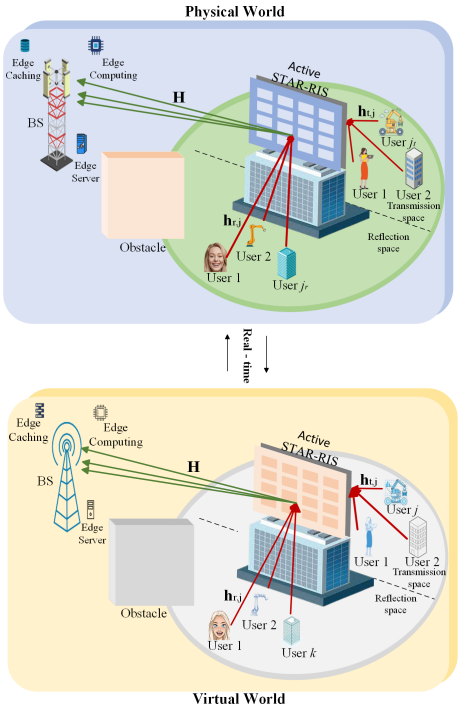

Fig. 1: An illustration of URLLC- active STAR-RIS DT system.

connectivity and robustness in scenarios where traditional communication methods falter due to high interference and latency issues. In Fig. 1, we also show its DT[1], which is linked to the physical network through a real-time connection. This system model specifically addresses the needs of industrial IoT applications such as automated manufacturing, where reliable, real-time data transmission is critical for machine-to-machine

[1] DTs operate in the cloud, not on devices, enhancing network management and operational efficiency without increasing device load. They provide real-time simulations for preemptive problem-solving and resource optimization, crucial for the sustainability and effectiveness of these IoT systems.

TABLE II: Table of notations.

Index		**Meaning**
j_r	$\in$	$\{1, 2, ..., J_r\}$ (index of j-th UN in the reflection space)
j_t	$\in$	$\{1, 2, ..., J_t\}$ (index of j-th UN in the transmission space)
j	$\in$	$\{1, 2, ..., J\}$ (index of all j-th UN)
n	$\in$	$\{1, 2, ..., N\}$ (index of n-th RIS element)
Notation		**Meaning**
$\mathbf{h}_{r,j}$	$\triangleq$	the reflection channel gain IoT UNs j^{th} to the active STAR-RIS
$\mathbf{h}_{t,j}$	$\triangleq$	the transmission channel gain IoT UNs j^{th} to the active STAR-RIS
$\mathbf{H}$	$\triangleq$	the channel gain from the active STAR-RIS to the BS
$\mathbf{h}_{ek}$	$\triangleq$	channel estimation error (CEE)
$\mathbf{\Psi}_r$	$\triangleq$	the reflection matrices for the active STAR-RIS
$\mathbf{\Psi}_t$	$\triangleq$	the transmission matrices for the active STAR-RIS
$\mathbf{A}_r$	$\triangleq$	the equivalent reflection amplitude matrix
$\mathbf{A}_t$	$\triangleq$	the equivalent transmission amplitude matrix
$\mathbf{B}_r$	$\triangleq$	the reflection amplitude coefficients
$\mathbf{B}_t$	$\triangleq$	the transmission amplitude coefficients
$\tilde{\mathbf{A}}$	$\triangleq$	the amplification matrix of the active STAR-RIS
$\mathcal{CN}(m, \sigma^2)$	$\triangleq$	Complex Gaussian distribution with mean m and variance σ^2
γ_j	$\triangleq$	SNR of the j^{th} IoT UN
R_j	$\triangleq$	rate of the j^{th} IoT UN
p_j	$\triangleq$	the transmit power of the j^{th} IoT UN
B	$\triangleq$	the total bandwidth of the system
b_j	$\triangleq$	the allocated bandwidth coefficient of the j^{th} IoT UN
V_j	$\triangleq$	the channel dispersion
f_j^{un}	$\triangleq$	the estimated processing rate of the j^{th} IoT UN
$\hat{f}_j^{un}$	$\triangleq$	the error in the processing rate estimation
β_j	$\triangleq$	proportion of the tasks
$\mathbf{w}_j$	$\triangleq$	the transmit beamforming vector
J_j	$\triangleq$	the task at the j^{th} IoT UN
D_j	$\triangleq$	data size
C_j	$\triangleq$	cycles for computation
δ	$\triangleq$	the transmission time interval
T	$\triangleq$	the cumulative latency
E	$\triangleq$	energy consumption
ξ	$\triangleq$	the energy conversion efficiency
σ_0^2	$\triangleq$	variance of AWGN at active STAR-RIS
σ_b^2	$\triangleq$	variance of AWGN at BS
σ_{ek}^2	$\triangleq$	variance of AWGN at CEE
$\lvert \cdot \rvert$	$\triangleq$	absolute value
$\lvert\lvert \cdot \rvert\rvert$	$\triangleq$	the norm of a value
$Q(\cdot)$	$\triangleq$	complementary cumulative distribution function (CCDF) of the standard normal distribution
$Q^{-1}(\cdot)$	$\triangleq$	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

communications and process automation. The inclusion of small cells ensures high-quality connections are maintained throughout the facility, enhancing the efficacy of the DT and active STAR-RIS in managing complex communication dynamics and ensuring URLLC[2].

There are J IoT UNs present in the real environment. The IoT UNs situated in the reflection space and the transmission space are denoted by the set of $\mathcal{J}_r \triangleq \{1, 2, ..., J_r\}$ and $\mathcal{J}_t \triangleq \{1, 2, ..., J_t\}$, respectively, with $J_r + J_t = J$. The single-antenna IoT UNs communicate with a M-antenna macro BS during the finite block-length transmission for task offloading. The active STAR-RIS has N elements. Hence, the reflection and transmission matrices for the active STAR-RIS are defined as $\mathbf{\Psi}_r \triangleq \mathbf{B}_r\tilde{\mathbf{A}}\mathbf{\Phi}_r = \mathrm{diag}\left(\boldsymbol{\psi}_r\right) \in \mathbb{C}^{N\times N}$ and $\mathbf{\Psi}_t \triangleq \mathbf{B}_t\tilde{\mathbf{A}}\mathbf{\Phi}_t = \mathrm{diag}\left(\boldsymbol{\psi}_t\right) \in \mathbb{C}^{N\times N}$, respectively, where $\mathbf{\Phi}_r \triangleq \mathrm{diag}\left(\phi_r\right) = \mathrm{diag}(\left[e^{j\varphi_{r,1}}, \ldots, e^{j\varphi_{r,N}}\right]^T) \in \mathbb{C}^{N\times N}$ denotes the reflection phase-shift matrix and $\mathbf{\Phi}_t \triangleq \mathrm{diag}\left(\boldsymbol{\phi}_t\right) = \mathrm{diag}(\left[e^{j\varphi_{t,1}}, \ldots, e^{j\varphi_{t,N}}\right]^T) \in \mathbb{C}^{N\times N}$ denotes the transmission phase-shift matrix [15]. Here, $\mathbf{B}_r \triangleq \mathrm{diag}\left(\boldsymbol{\beta}_r\right) = \mathrm{diag}([\beta_1^r, \ldots, \beta_N^r]^T) \in \mathbb{C}^{N\times N}$ and $\mathbf{B}_t \triangleq \mathrm{diag}\left(\boldsymbol{\beta}_t\right) = \mathrm{diag}([\beta_1^t, \ldots, \beta_N^t]^T) \in \mathbb{C}^{N\times N}$ denote the reflection and transmission amplitude coefficients, respectively, while $\tilde{\mathbf{A}} \triangleq \mathrm{diag}(\tilde{\boldsymbol{a}}) = \mathrm{diag}([\tilde{a}_1, \ldots, \tilde{a}_N]^T) \in \mathbb{C}^{N\times N}$ is the amplification matrix of the active STAR-RIS.

We define $\mathbf{A}_r \triangleq \mathbf{B}_r\tilde{\mathbf{A}} \in \mathbb{C}^{N\times N}$ and $\mathbf{A}_t \triangleq \mathbf{B}_t\tilde{\mathbf{A}} \in \mathbb{C}^{N\times N}$ as the equivalent reflection and transmission amplitude

[2]By incorporating DT into our URLLC IoT system, we can create virtual replicas of physical entities for real-time monitoring, analysis, and prediction [7], [8], [38]. This integration enables informed decision-making, improves resource management, and enhances operational efficiency. The system's real-time analytics and low-latency processing amplify scalability, flexibility, and cost-efficiency. It also reduces network congestion, strengthens data privacy and security, and enhances user experiences. Offline functionality ensures continuity in limited connectivity environments, making DT with URLLC IoT system a compelling solution for process optimization and innovation [8], [39]. Overall, DT provides a scalable, immersive, and efficient platform for the URLLC IoT system by enabling real-time interaction between users and digital objects and reducing the reliance on central nodes [7]–[9], [34], [38]–[42].

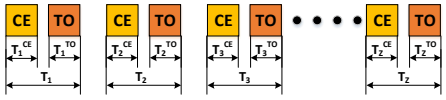

Fig. 2: Illustration of time-frame.

matrix, respectively. Thus, we have $\mathbf{A}_r \triangleq \text{diag}(\boldsymbol{a}_r) = \text{diag}([a_{r,1}, \ldots, a_{r,N}]^T) = \text{diag}([\tilde{a}_1\beta_1^r, \ldots, \tilde{a}_N\beta_N^r]^T)$, $\mathbf{A}_t \triangleq \text{diag}(\boldsymbol{a}_t) = \text{diag}([a_{t,1}, \ldots, a_{t,N}]^T) = \text{diag}([\tilde{a}_1\beta_1^t, \ldots, \tilde{a}_N\beta_N^r]^T)$ and their corresponding reflection and transmission vectors can be written as $\boldsymbol{\psi}_r = \text{diag}([a_{r,1}\phi_{r,1}, \ldots, a_{r,N}\phi_{r,N}]^T)$ and $\boldsymbol{\psi}_t = \text{diag}([a_{t,1}\phi_{t,1}, \ldots, a_{t,N}\phi_{t,N}]^T)$, respectively. We define $\mathbf{H} \in \mathbb{C}^{M\times N}$, $\mathbf{h}_{r,j} \in \mathbb{C}^{N\times 1}$ as the reflection channel gain from the active STAR-RIS to the BS and IoT UNs j^{th} to the active STAR-RIS, and $\mathbf{h}_{t,j} \in \mathbb{C}^{N\times 1}$ as the transmission channel gain from the active STAR-RIS to the BS and IoT UNs j^{th} to the active STAR-RIS, respectively.

A. Channel estimation

Fig. 2 illustrates the time-frame model used in our study. The diagram labels "CE" and "TO" represent the time frames allocated for channel estimation and task offloading, respectively. Each independent time-frame, denoted as $\mathrm{T_z}$ where $\mathcal{Z} = \{1, \ldots, Z\}$, is divided into two sub-time-frames: $\mathrm{T_z}^{\mathrm{CE}}$ for channel estimation and $\mathrm{T_z}^{\mathrm{TO}}$ for task offloading. In this work, we do not consider the channel estimation process in the problem formulation, as it represents a different aspect of research that is not our concern for the present work. We simply assume that channel estimation is completed immediately before task offloading in each timeframe. Relevant nodes estimate channel using standard techniques, as detailed in [43]. This work in [43] provides a thorough overview and various methods for estimating channels in RIS systems. While our primary focus is on the task offloading scheme, utilizing prior channel estimation knowledge, it is important to note that the CSI may contain errors. Therefore, we model the imperfections in CSI as discussed in the preceding subsection.

1) Imperfect CSI modeling: Considering an imperfect CSI[3] scenario, the actual reflection and transmission channels can be represented in terms of their estimates as follows:

$$\mathbf{h}_{j,r} = \hat{\mathbf{h}}_{j,r} + \mathbf{h}_{e,r}, \text{ where } \hat{\mathbf{h}}_{j,r} = \mathbf{H}\boldsymbol{\Psi}\mathbf{h}_{r,j}, \quad (1)$$

$$\mathbf{h}_{j,t} = \hat{\mathbf{h}}_{j,t} + \mathbf{h}_{e,t}, \text{ where } \hat{\mathbf{h}}_{j,t} = \mathbf{H}\boldsymbol{\Psi}\mathbf{h}_{t,j}. \quad (2)$$

Here, $\mathbf{h}_{j,r} \in \mathbb{C}^{M\times 1}$, and $\mathbf{h}_{j,t} \in \mathbb{C}^{M\times 1}$ represent the actual reflection and transmission channels, respectively. The terms $\hat{\mathbf{h}}_{j,r}$ and $\hat{\mathbf{h}}_{j,t}$ are the estimated CSI for the reflection and transmission channels, respectively. $\mathbf{h}_{e,r} \in \mathbb{C}^{M\times 1}$ and $\mathbf{h}_{e,t} \in \mathbb{C}^{M\times 1}$ denotes the estimation error in the reflection and transmission channel, respectively.

The terms $\mathbf{h}_{e,t}(z) \in \mathbb{C}^{M\times 1}$ in (1) and $\mathbf{h}_{e,r}(z) \in \mathbb{C}^{M\times 1}$ in (2) are the CSI estimation errors. It is characterized by the set as follows [46]:

$$\mathcal{T}_\epsilon = \left\{\mathbf{h}_{e,t} \in \mathbb{C}^{M\times 1} : \|\mathbf{h}_{e,t}\| \leq \tau_{e,t},\, e,t = 1, \ldots, M\right\}, \quad (3)$$

$$\mathcal{T}_\epsilon = \left\{\mathbf{h}_{e,r} \in \mathbb{C}^{M\times 1} : \|\mathbf{h}_{e,r}\| \leq \tau_{e,r},\, e,r = 1, \ldots, M\right\}, \quad (4)$$

where $\|\cdot\|$ denotes the norm. Within the set, entries of $\mathbf{h}_{e,t}$ and $\mathbf{h}_{e,r}$ are independent and identically distributed (i.i.d) and assumed to have zero mean with variance $\sigma_{e,t}^2$ and $\sigma_{e,r}^2$, respectively. For the present work, we consider uniformly distributed bounded CSI uncertainty $\sigma_{e,t}^2 = \sigma_{e,r}^2 = \{0.05, 0.10\}$, as discussed in [46], [47].

B. Signal model

The received signal at the BS from the IoT UNs in the reflection space is given as

$$y_j = \mathbf{w}_j(\mathbf{h}_{j,r}\mathbf{x}_j + \mathbf{H}\boldsymbol{\Psi}_r\mathbf{z}_0 + \mathbf{n}_b), j \in \mathcal{J}_r, \quad (5)$$

where $\mathbf{w}_j$ is the active receive beamformer at the BS, $\mathbf{x}_j$ is the data symbol sent by the k^{th} IoT UN, $\mathbf{z}_0 \sim \mathcal{CN}(0, \sigma_0^2\mathbf{I}_N)$ and $\mathbf{n}_b \sim \mathcal{CN}(0, \sigma_b^2\mathbf{I}_M)$ represent the dynamic noise at the active STAR-RIS and the additive white Gaussian noise (AWGN) added at the BS, respectively. The SNR of the j^{th} IoT UN in the reflection space is given by

$$\gamma_j = \frac{|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,r}|^2 p_j}{||\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_r||^2\sigma_0^2 + ||\mathbf{w}_j^H||^2 B b_j N_0}, j \in \mathcal{J}_r, \quad (6)$$

where p_j is the transmit power of the j^{th} IoT UN, B is the total bandwidth of the system, and N_0 represents the effective noise, which is a combination of the noise added at the BS and CEE, given by $N_0 = \sigma_b^2 + \sigma_{ek}^2$. Similarly, the received signal at the BS from the IoT UNs within the transmission space can be calculated as

$$y_j = (\hat{\mathbf{h}}_{j,t} + \mathbf{h}_{ek})\mathbf{x}_j + \mathbf{H}\boldsymbol{\Psi}_t\mathbf{z}_0 + \mathbf{n}_b, j \in \mathcal{J}_t. \quad (7)$$

The SNR of the j^{th} IoT UN in the transmission space can be expressed as

$$\gamma_j = \frac{|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,t}|^2 p_j}{||\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_t||^2\sigma_0^2 + ||\mathbf{w}_j^H||^2 B b_j N_0}, j \in \mathcal{J}_t. \quad (8)$$

C. DT-assisted active STAR-RIS communication model

As there is no direct path available due to the obstacles in the path between the IoT UNs and the BS, all the IoT UNs transmit their signal to the BS with the aid of active STAR-RIS. The uplink rate expression for the j^{th} IoT UN is given as [34], [48]

$$R_j = \frac{B}{\ln 2}\left[b_j \ln(1+\gamma_j) - \sqrt{\frac{b_j V_j}{\delta B}}Q^{-1}(\varepsilon_j)\right], \quad (9)$$

where b_j is the allocated bandwidth coefficient of the j^{th} IoT UN, δ is the transmission time interval, $V_j = 1 - [1+\gamma_j]^{-2}$

[3]The channel estimation for active Simultaneous Transmit and Reflect Reconfigurable Intelligent Surface (STAR-RIS) requires tailored approaches due to its signal amplification capabilities and additional noise [44]. Aggregated channel estimation methods can reduce overhead by leveraging aggregated channels for data processing. Two-timescale designs, such as iterative methods based on the proximal gradient amplification method (PGAM), update optimization parameters simultaneously per iteration, accommodating the active components of STAR-RIS [45]. Hybrid channel estimation combines active and passive techniques, where active sensors estimate separate links and passive patterns estimate cascade links. Statistical CSI may be used for passive beamforming design, particularly in spatially-correlated channel scenarios. These methods are adapted to the unique characteristics of active STAR-RIS, ensuring effective CSI acquisition for optimal network performance [44].

is the channel dispersion and Q^{-1} (.) is the inverse function, $Q(\varepsilon_j) = \frac{1}{\sqrt{2\pi}} \int_{\varepsilon_j}^{\infty} e^{\left(\frac{-t^2}{2}\right)} dt$. The uplink transmission latency is given as $T_j^{\text{co}} = \frac{D_j}{R_j}$, where D_j is data size (bits).

D. DT-assisted active STAR-RIS computation model

The j^{th} IoT UN locally executes β_j proportion of the tasks and offloads the $(1-\beta_j)$ proportion of tasks to the ECS. A tuple $J_j = (D_j, C_j, T_j^{max})$ represents the task at the j^{th} IoT UN, where C_j and T_j^{max} are the required cycles for computation and the maximum task latency, respectively.

We model the DT service for local processing as $DT_j^{un} = (f_j^{un}, \hat{f}_j^{un})$, where f_j^{un} denotes the estimated processing rate of the j^{th} IoT UNs and $\hat{f}_j^{un}$ represents the error in the processing rate estimation.

The processing latency of node P is expressed as

$$T_j^P = \widetilde{T}_j^P + T_{ek}^P = \zeta_j C_j / f_{ek}^P, \tag{10}$$

where $P \in \{un, ecs\}$, $\zeta_j = \beta_j$ for $P = un$ and $\zeta_j = 1 - \beta_j$ for $P = ecs$. Here, $\widetilde{T}_j^P = \zeta_j C_j / f_j^P$ is the estimated processing latency, $T_{ek}^P = \frac{\zeta_j C_j \hat{f}_j^P}{\left[f_j^P (f_{ek}^P)\right]}$ is the latency error and $f_{ek}^P = f_j^P - \hat{f}_j^P$ is the processing rate error.

E. Energy and Latency computation

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

$$E_j^{cp} = \beta_j \theta C_j \left(f_j^{un} - \hat{f}_j^{un}\right)^2 /2, \tag{11}$$

$$E_j^{cm} = (1-\beta_j)\, p_j D_j / R_j, \tag{12}$$

$$E_j^{tot} = (1-\mu_j)\left[E_j^{cp} + E_j^{cm}\right], \tag{13}$$

where E_j^{cp}, E_j^{cm}, E_j^{tot} and θ are the energy for computation, the energy for communication, the total energy consumption and its power parameter, respectively. Here, $\boldsymbol{\mu} \triangleq \{\mu_j\} | \mu_j \in \{0,1\}, \forall j$, which represents the IoT UNs affiliation with the ECS, i.e., when $\mu_j = 1$, there exists a connection between IoT UNs and ECS and thereby the task J_j is cached at the ECS which is offloaded from the IoT UNs, and when $\mu_j = 0$, there exists no connection between IoT UNs and ECS.

The cumulative latency considering MEC is expressed as

$$T_j^{tot} = \frac{\mu_j C_j}{f_{ek}^{ecs}} + (1-\mu_j) \times \left[T_j^{un} + T_j^{\text{co}} + T_j^{ecs}\right]. \tag{14}$$

F. Optimization Problem formulation

In our system, a crucial small CPU at the BS aggregates and processes information from IoT nodes, MEC-enabled BS, and active STAR-RIS. This CPU, integral to the BS, uses a control channel to gather data for decision-making, as detailed in equation (15). After processing, it executes an optimization algorithm to minimize total task offloading latency, accounting for channel condition variations with imperfect CSI. This architecture ensures practical, efficient real-world application management.

Our objective is to minimize the total latency of IoT UNs by dealing with the optimization of the reflection and transmission beamforming matrices at the active STAR-RIS, allocated bandwidth, offloading proportions, transmit power, estimated processing rates at IoT UNs and ECS, the energy consumption of IoT UNs and MEC capacity at ECS. Thus, the optimization problem can be formulated as follows

$$\min_{\mathbf{w}_j, \beta_j, \mu_j, b_j, p_j, \boldsymbol{\Psi}_r, \boldsymbol{\Psi}_t, f_j^{\text{un}}, f_j^{\text{ecs}}} \sum_{j=1}^{J} T_j^{\text{tot}} \tag{15a}$$

$$\text{s.t.} \quad T_j^{\text{tot}} \le T_j^{\max}, \forall j, \tag{15b}$$

$$\sum_{j=1}^{J} \left[\mu_j f_j^{\text{ecs}} + (1-\mu_j)(1-\beta_j) f_j^{\text{ecs}}\right] \le F_{\max}^{\text{ecs}}, \tag{15c}$$

$$E_j^{\text{tot}} \le E_j^{\max}, \forall j, \tag{15d}$$

$$R_j \ge R_{\min}, \forall j, \tag{15e}$$

$$\sum_{j=1}^{J} b_j \le 1, \forall j, \tag{15f}$$

$$\sum_{j=1}^{J} \mu_j D_j \le S_{\max}^{\text{ecs}}, \tag{15g}$$

$$\mathbf{p} \in \mathscr{P}, \boldsymbol{\beta} \in \mathscr{B}, \mathbf{f} \in \mathscr{F}, \tag{15h}$$

$$\|\mathbf{w}_j\| = 1, \forall j, \tag{15i}$$

$$\left\|\mathbf{w}_j^H \mathbf{H} \boldsymbol{\Psi}_t\right\|^2 \sigma_0^2 + \left\|\mathbf{w}_j^H \mathbf{H} \boldsymbol{\Psi}_r\right\|^2 \sigma_0^2 \le P_{max}^{RIS}, \tag{15j}$$

$$a_{r,n} \ge 0, a_{t,n} \ge 0, \forall n, \tag{15k}$$

$$|\phi_{r,n}| = 1, |\phi_{t,n}| = 1, \forall n, \tag{15l}$$

where $\mathscr{P} \triangleq \left\{p_j, \forall j \mid 0 \le p_j \le P_j^{max}, \forall j\right\}$, $\mathscr{B} \triangleq \{\beta_j, \forall j \mid 0 \le \beta_j \le 1, \forall j\}$, $\mathscr{F} \triangleq \left\{\mathbf{f} = \left\{f_j^{\text{un}}, f_j^{\text{ecs}}\right\}, \forall j | 0 \le f_j^{\text{un}} \le F_{\max}^{\text{un}}, \forall j; 0 \le f_j^{\text{ecs}} \le F_{\max}^{\text{ecs}} \forall j\right\}$ are the constraints associated with the UL power, offloading decisions, and processing rates, respectively. The constraints related to the upper limit of the latency, computation capacity, and energy of the IoT UNs are given by (15b), (15c), and (15d), respectively. Moreover, (15e), (15f), (15g), and (15i) define the constraints corresponding to the minimum transmission rate, bandwidth allocation of IoT UNs, the caching capacity of the ECS, and the transmit beamforming, respectively. The constraints related to the maximum power, the amplitude coefficient and the unit modulus phase at the active STAR-RIS are given by (15j), (15k) and (15l), respectively. Here, (15) represents a non-convex optimization problem due to its intricate objective function (15a), which includes logarithmic and fractional terms, as well as the interdependencies between variables in both the objective function and constraints. Hence, to tackle this challenge, we propose an efficient AO algorithm.

III. Proposed Solution

The optimization challenge presented in (15) is characterized by a non-convex objective function and a complex interplay between continuous and integer variables, making it a highly intricate problem to solve directly. To address this, we introduce an AO algorithm that simplifies the problem by isolating and sequentially optimizing individual variables, keeping all others fixed. This methodical approach decomposes the

original problem into five manageable subproblems, namely, beamforming design, caching policy optimization, offloading policy optimization, joint computation and communication, and active STAR-RIS beamforming optimization. The more detailed analysis of each subproblem is provided as follows

A. Beamforming Design

In this subsection, we focus on optimising beamforming vectors, denoted as $\mathbf{w}_j$, which are crucial for directing the signal power toward intended IoT UNs while minimizing interference. The objective is to minimize the total transmission power, T_j^{tot}, subject to a set of constraints that ensure reliable communication. The optimization subproblem for the beamforming design can be formulated as:

$$\min_{\mathbf{w}_j|\boldsymbol{\mu}^{(i)},\boldsymbol{\beta}^{(i)},\mathbf{f}^{(i)},\mathbf{b}^{(i)},\mathbf{p}^{(i)},\boldsymbol{\Psi_r}^{(i)},\boldsymbol{\Psi_t}^{(i)}} \sum_{j=1}^{J} T_j^{\text{tot}}\left(\mathbf{w}_j\right) \tag{16a}$$

$$\text{s.t.} \quad (15\text{b}), (15\text{e}), (15\text{i}), (15\text{j}), \tag{16b}$$

where $\boldsymbol{\mu} \in \mu_j$, $f \in \{f_j^{un}, f_j^{ecs}\}$, $\boldsymbol{\beta} \in \beta_j$, $\mathbf{b} \in b_j$, and $\mathbf{p} \in p_j, \forall j$.

To achieve the maximum SNR for each IoT UN, we employ the linear minimum mean square error (MMSE) receiver, which is recognized for its efficiency in balancing signal enhancement and interference suppression. The MMSE-based equalizer for IoT UNs is given by [49]

$$\mathbf{w}_j^* = \frac{\left(\mathbf{h}_j\mathbf{h}_j^H p_j + \mathbf{H}\boldsymbol{\Psi}\boldsymbol{\Psi}^H\mathbf{H}^H\sigma_0^2 + \sigma_b^2\mathbf{I}_M\right)^{-1}\mathbf{h}_j\sqrt{p_j}}{\left\|\left(\mathbf{h}_j\mathbf{h}_j^H p_j + \mathbf{H}\boldsymbol{\Psi}\boldsymbol{\Psi}^H\mathbf{H}^H\sigma_0^2 + \sigma_b^2\mathbf{I}_M\right)^{-1}\mathbf{h}_j\sqrt{p_j}\right\|}. \tag{17}$$

B. Caching Policy Optimization

To determine the optimal content to be cached at the edge nodes, this subsection aims to optimize the caching policy by determining the next iteration point for the caching variables, $\boldsymbol{\mu}^{(i+1)}$, while keeping all other system constraints constant. The optimization subproblem for the caching variables is formulated as follows:

$$\min_{\mu_j|,\mathbf{w}_j^{(i+1)},\boldsymbol{\beta}^{(i)},\mathbf{f}^{(i)},\mathbf{b}^{(i)},\mathbf{p}^{(i)},\boldsymbol{\Psi_r}^{(i)},\boldsymbol{\Psi_t}^{(i)}} \sum_{j=1}^{J} T_j^{\text{tot}}\left(\mu_j\right) \tag{18a}$$

$$\text{s.t.} \quad (15\text{b}), (15\text{c}), (15\text{d}), (15\text{g}). \tag{18b}$$

Given that μ_j are integer variables, the problem is inherently non-convex. To tackle this challenge, we define a new variable $t_j^s = T_j^{un} + T_j^{co} + T_j^{ecs}$ for each task j, which represents the total time saved by caching task j. We then sort these values in descending order and prioritize caching the tasks with the greatest time savings. This heuristic approach continues until adding another task would violate the system's capacity constraint. By employing this method, we can efficiently approximate the optimal caching policy $\boldsymbol{\mu}$ at each iteration with a reduced number of constraint checks, which is less than the total number of tasks J, thus simplifying the optimization process.

C. Offloading Policy Optimization

This subsection is dedicated to developing the most effective strategy for offloading computational tasks from IoT UNs to the ECS. The focus is on optimizing the offloading policy by determining the optimal offloading decisions, $\boldsymbol{\beta}^{(i+1)}$, for the next iteration. The optimization problem is structured to minimize the total time spent on task execution across all tasks, denoted by $T_j^{\text{tot}}\left(\beta_j\right)$, subject to a set of system constraints. The optimization subproblem is formulated as:

$$\min_{\beta_j|\mathbf{w}_j^{(i+1)},\boldsymbol{\mu}^{(i+1)},\mathbf{f}^{(i)},\mathbf{b}^{(i)},\mathbf{p}^{(i)},\boldsymbol{\Psi_r}^{(i)},\boldsymbol{\Psi_t}^{(i)}} \sum_{j=1}^{J} T_j^{\text{tot}}\left(\beta_j\right) \tag{19a}$$

$$\text{s.t.} \quad (15\text{b}), (15\text{c}), (15\text{d}), (15\text{h}). \tag{19b}$$

Since the constraints of this problem are linear, the problem is convex. Hence, it can be solved efficiently using standard convex optimization techniques. One such technique is using the CVX toolbox, which is designed for solving convex optimization problems [8].

D. Joint Communication and Computation Optimization

In this subsection, our objective is to determine the subsequent iteration points $\mathbf{f}^{(i+1)}, \mathbf{b}^{(i+1)}, \mathbf{p}^{(i+1)}$ by fixing the values of $\boldsymbol{\mu}^{(i+1)}$ and $\boldsymbol{\beta}^{(i+1)}$. Utilizing the approximation $V_j \approx 1$ under the assumption of high received SNR [8], R_j is given as

$$\begin{aligned} R_j &\approx \frac{B}{\ln 2}\left[b_j \ln\left(1+\gamma_j\right) - \sqrt{\frac{b_j}{\delta B}}Q^{-1}(\varepsilon_j)\right] \\ &\triangleq \frac{B}{\ln 2}\left[\mathbb{G}_j - \mathbb{B}_j\right], \end{aligned} \tag{20}$$

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

$$\mathbb{G}_j^{(i)} = z\ln\left(1+\frac{\hat{x}}{\hat{y}}\right) + x\left(\frac{\hat{z}}{\hat{x}+\hat{y}}\right) - y\left(\frac{\hat{x}\hat{z}}{\hat{y}(\hat{x}+\hat{y})}\right). \tag{21}$$

Moreover, by using the inequality $\sqrt{z} \leq \frac{\sqrt{\hat{z}}}{2} + \frac{z}{2\sqrt{\hat{z}}}$, we can approximate $\mathbb{B}_j^{(i)}$ as $\mathbb{B}_j \leq \frac{Q^{-1}(\varepsilon_j)}{\sqrt{\delta B}}(\frac{\sqrt{\hat{z}}}{2} + \frac{z}{2\sqrt{\hat{z}}}) \triangleq \mathbb{B}_j^{(i)}$, where $z = b_j$, $x = |\mathbf{w}_j^H\hat{\mathbf{h}}_j|^2 p_j$, $y = ||\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}||^2\sigma_0^2 + ||\mathbf{w}_j^H||^2 B b_j N_0$, $\hat{z} = \hat{b}_j$, $\hat{x} = |\mathbf{w}_j^H\hat{\mathbf{h}}_j|^2\hat{p}_j$, and $\hat{y} = ||\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}||^2\sigma_0^2 + ||\mathbf{w}_j^H||^2 B\hat{b}_j N_0$. Thus, the rate expression is given by a lower bound as $R_j \geq R_j^{(i)} \triangleq \frac{B}{\ln 2}\left[\mathbb{G}_j^{(i)} - \mathbb{B}_j^{(i)}\right]$. Now, the approximate expression of constraint (15f) is $R_j^{(i)} \geq R_{min}, \forall j$. Now, we introduce a new variable $\boldsymbol{\tau_j} \triangleq \{\tau_j\}, \forall j$ to transform the non-convex rate constraint (i.e., (15e)) into a convex one as follows. Consider $1/R_j \leq \tau_j, \forall j$ and thus, we can reformulate the constraint (15e) as

$$1/R_j^{(i)} \leq \tau_j, \tag{22}$$

$$\begin{aligned} &(1-\mu_j^{(i+1)})\left[\frac{\theta}{2}\beta_j^{(i+1)}C_j(f_{ek}^{un})^2\right. \\ &\left.+(1-\beta_j^{(i+1)})p_j\tau_j\right] \leq E_j^{max}, \forall j. \end{aligned} \tag{23}$$

Now, constraint (22) is convex. However, we observe that (23) is still non-convex. Hence, we approximate the constraint using the inequality given in [8] as follows

$$\left(1-\mu_j^{(i+1)}\right)\left[\frac{\theta\beta_j^{(i+1)}C_j(f_{ek}^{un})^2}{2} + \frac{\left(1-\mu_j^{(i+1)}\right)}{2}\left(\frac{\tau_j^{(i)}}{p_j^{(i)}}p_j^2+\frac{p_j^{(i)}}{\tau_j^{(i)}}\tau_j^2\right)\right] \leq E_j^{\max}, \forall j. \quad (24)$$

Then, the non-convex objective function (15a) can be approximately represented as follows

$$T_j^{tot} \leq (1-\mu_j^{(i+1)})\left[\frac{\beta_j^{(i+1)}C_j}{f_{ek}^{un}} + D_j\tau_j + \frac{(1-\mu_j^{(i+1)})C_j}{f_{ek}^{ecs}}\right] + \frac{\mu_j^{(i+1)}C_j}{f_{ek}^{ecs}} \triangleq \acute{T}_j^{tot}. \quad (25)$$

Finally, we can reformulate subproblem (19) as [8]

$$\min_{\mathbf{b},\mathbf{p},\mathbf{f}|\mathbf{w}_j^{(i+1)},\boldsymbol{\mu}^{(i+1)},\boldsymbol{\beta}^{(i+1)},\boldsymbol{\Psi_r}^{(i)},\boldsymbol{\Psi_t}^{(i)}} \sum_{j=1}^{J}\acute{T}_j^{tot}, \forall j \quad (26a)$$

$$\text{s.t.} \quad \acute{T}_j^{tot} \leq T_j^{max}, \quad (26b)$$

$$(15c), (15e), (15f), (15h), (22), (24). \quad (26c)$$

This subproblem is solved using the AO problem as described in step 6 of Algorithm 1.

E. Active STAR-RIS Beamforming Optimization

This section focuses on optimising the $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_t$ for the active STAR-RIS. To address this challenge, we consider the rate expression as provided below:

$$\max_{\substack{\boldsymbol{\Psi}_r,\boldsymbol{\Psi}_t|\mathbf{b}^{(i+1)},\mathbf{p}^{(i+1)},\mathbf{f}^{(i+1)}\\ \mathbf{w}_j^{(i+1)},\boldsymbol{\mu}^{(i+1)},\boldsymbol{\beta}^{(i+1)}}} R_j, \ \ \forall j \quad (27a)$$

$$\text{s.t.} \quad (15j), (15k), (15l), \quad (27b)$$

To convert the objective function (27a) into a more tractable form, we suggest a fractional programming (FP) algorithm. Following this, by introducing an auxiliary variable and applying the Lagrangian dual transform $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_J]^T$, we can reframe the initial objective function (27) as

$$\begin{aligned}\tilde{R}_j =& \sum_{j=1}^{J}(1+\gamma_j) - \sum_{j=1}^{J}\gamma_j \\ &+ \sum_{j\in\mathcal{J}_r}\frac{(1+\gamma_j)\left|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,r}\right|^2}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_r\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0} \\ &+ \sum_{j\in\mathcal{J}_t}\frac{(1+\gamma_j)\left|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,t}\right|^2}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_t\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0}. \end{aligned} \quad (28)$$

Subsequently, we proceed by applying the quadratic transform to the remaining two fractional terms in (29) by introducing another auxiliary variable $\boldsymbol{\rho} = [\rho_1, ..., \rho_J]^T$. This allows us to further reconfigure (28) into the form presented as

$$\begin{aligned} f(\gamma,\boldsymbol{\Psi}_r,\boldsymbol{\Psi}_t,\rho) =& \sum_{j\in\mathcal{J}_r}(2\sqrt{1+\gamma_j}\Re\left\{\rho_j^*\mathbf{w}_j^H\hat{\mathbf{h}}_{j,r}\right\} \\ &- |\rho_j|^2\left(\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_r\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0\right)) \\ &+ \sum_{j\in\mathcal{J}_r}(2\sqrt{1+\gamma_j}\Re\left\{\rho_j^*\mathbf{w}_j^H\hat{\mathbf{h}}_{j,t}\right\} \\ &- |\rho_j|^2\left(\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_t\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0\right)). \end{aligned} \quad (29)$$

Now, we obtain optimal auxiliary variables γ and ρ by letting $\partial\tilde{R}_j/\partial\gamma_j$ and $\partial f/\partial\rho_k$ to 0,

$$\gamma_j^{opt} = \begin{cases} \frac{\left|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,r}\right|^2}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_r\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0}, & j\in\mathcal{J}_r, \\ \frac{\left|\mathbf{w}_j^H\hat{\mathbf{h}}_{j,r}\right|^2}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_t\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0}, & j\in\mathcal{J}_t, \end{cases} \quad (30)$$

$$\rho_j^{opt} = \begin{cases} \frac{\sqrt{1+\gamma_j}\mathbf{w}_j^H\hat{\mathbf{h}}_{j,t}}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_r\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0}, & j\in\mathcal{J}_r, \\ \frac{\sqrt{1+\gamma_j}\mathbf{w}_j^H\hat{\mathbf{h}}_{j,t}}{\left\|\mathbf{w}_j^H\mathbf{H}\boldsymbol{\Psi}_t\right\|^2\sigma_0^2+\left\|\mathbf{w}_j^H\right\|^2 Bb_jN_0}, & j\in\mathcal{J}_t. \end{cases} \quad (31)$$

The active STAR-RIS beamforming is solved under unequal energy division (UED) mode. Each component of the active STAR-RIS possesses the ability to transmit and reflect incoming signals simultaneously, with different amplitudes and phases. For ease of notation, we denote $\mathbf{w}_j^H\hat{\mathbf{h}}_j \triangleq \mathbf{h}_{r,j}\,\mathrm{diag}(\mathbf{w}_j^H\mathbf{H})\psi, j\in\mathcal{J}_r$ and $\mathbf{w}_j^H\hat{\mathbf{h}}_j \triangleq \mathbf{h}_{t,j}\,\mathrm{diag}(\mathbf{w}_j^H\mathbf{H})\psi, j\in\mathcal{J}_t$. Similarly, we define

$$\mathbf{m}_r \triangleq \sum_{j\in\mathcal{J}_r}\mathrm{diag}(\mathbf{h}_{r,j})(2\sqrt{1+\gamma_j}\rho^*\mathbf{w}_j^H\mathbf{H}), \quad (32a)$$

$$\mathbf{M}_r \triangleq \sum_{j\in\mathcal{J}_r}|\rho_j|^2\sigma_0^2\,\mathrm{diag}(\mathbf{w}_j^H\mathbf{H}\odot(\mathbf{w}_j^H\mathbf{H})^*), \quad (32b)$$

$$\mathbf{m}_t \triangleq \sum_{j\in\mathcal{J}_t}\mathrm{diag}(\mathbf{h}_{t,j})(2\sqrt{1+\gamma_j}\rho^*\mathbf{w}_j^H\mathbf{H}), \quad (32c)$$

$$\mathbf{M}_t \triangleq \sum_{j\in\mathcal{J}_t}|\rho_j|^2\sigma_0^2\,\mathrm{diag}(\mathbf{w}_j^H\mathbf{H}\odot(\mathbf{w}_j^H\mathbf{H})^*), \quad (32d)$$

$$\mathbf{O} = \mathrm{diag}(\mathbf{w}_j^H\mathbf{H}\odot(\mathbf{w}_j^H\mathbf{H})^*). \quad (32e)$$

Now, we reformulate the subproblem (27) with respect to $\boldsymbol{\psi}_r$ and $\boldsymbol{\psi}_t$ as

$$\max_{\boldsymbol{\psi}_r,\boldsymbol{\psi}_t}\boldsymbol{\psi}_r^H\mathbf{M}_r\boldsymbol{\psi}_r - \Re\left\{\boldsymbol{\psi}_r^H\mathbf{m}_r\right\} + \boldsymbol{\psi}t^H\mathbf{M}_t\boldsymbol{\psi}_t - \Re\left\{\boldsymbol{\psi}_t^H\mathbf{m}_t\right\} \quad (33a)$$

$$\text{s.t.} \quad \boldsymbol{\psi}_r^H\mathbf{O}\boldsymbol{\psi}_r + \boldsymbol{\psi}_t^H\mathbf{O}\boldsymbol{\psi}_t \geq P_{max}^R. \quad (33b)$$

Here, (33) is a QCQP problem that can be addressed with the standard convex optimization algorithm. By considering the constraints (15k) and (15l), the amplification factor vector $\boldsymbol{a}_r^{opt} \in \mathbb{C}^{N\times N}, \boldsymbol{a}_t^{opt} \in \mathbb{C}^{N\times N}$ and the associated phase-shift matrix $\boldsymbol{\Phi}_r^{opt} \in \mathbb{C}^{N\times N}, \boldsymbol{\Phi}_t^{opt} \in \mathbb{C}^{N\times N}$ are given by

$$\boldsymbol{a}_r^{opt} = \mathrm{diag}(e^{(-j\arg(\boldsymbol{\psi}_r^{opt}))})\boldsymbol{\psi}_r^{opt}, \quad (34a)$$

$$\boldsymbol{a}_t^{opt} = \mathrm{diag}(e^{(-j\arg(\boldsymbol{\psi}_t^{opt}))})\boldsymbol{\psi}_t^{opt}, \quad (34b)$$

$$\boldsymbol{\Phi}_r^{opt} = \mathrm{diag}(e^{(j\arg(\boldsymbol{\psi}_r^{opt}))}), \quad (34c)$$

$$\boldsymbol{\Phi}_t^{opt} = \mathrm{diag}(e^{(j\arg(\boldsymbol{\psi}_t^{opt}))}). \quad (34d)$$

We can obtain the optimal $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_t$, respectively. Finally, we summarize the above procedures in **Algorithm 1**.

Algorithm 1 illustrates the proposed AO framework that enables the joint optimization of all variables considered. This framework integrates the joint optimization solutions of the

Algorithm 1 AO-based algorithm for solving (15)

1: **Initialize:**
- Set i = 0 and $I^{max} = 50$;
- Set randomly choose initial feasible points $\mathbf{w}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\mathbf{b}^{(0)}$, $\mathbf{p}^{(0)}$, $\mathbf{f}^{(0)}$, $\boldsymbol{\Psi}_r^{(0)}$, and $\boldsymbol{\Psi}_t^{(0)}$;
- Set the tolerance $\epsilon_k = 10^{-3}$;

2: **repeat**

3: Solve (16) for given $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\beta}^{(i)}$,$\mathbf{f}^{(i)}$,$\mathbf{b}^{(i)}$,$\mathbf{p}^{(i)}$, $\boldsymbol{\Psi}_r^{(i)}$ and $\boldsymbol{\Psi}_t^{(i)}$ by following subsection **III-A** to find the best solution of $\mathbf{w}^\star$ and then update $\mathbf{w}^{(i+1)} = \mathbf{w}^\star$;

4: Solve (18) for given $\mathbf{w}^{(i)}$, $\boldsymbol{\beta}^{(i)}$,$\mathbf{f}^{(i)}$,$\mathbf{b}^{(i)}$,$\mathbf{p}^{(i)}$,$\boldsymbol{\Psi}_r^{(i)}$, and $\boldsymbol{\Psi}_r^{(i)}$; by following subsection **III-B** to find the best solution of $\boldsymbol{\mu}^\star$ and then update $\boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^\star$;

5: Solve (19) using $\mathbf{w}^{(i)}$, $\boldsymbol{\mu}^{(i)}$, $\mathbf{f}^{(i)}$,$\mathbf{b}^{(i)}$,$\mathbf{p}^{(i)}$, $\boldsymbol{\Psi}_r^{(i)}$, and $\boldsymbol{\Psi}_t^{(i)}$ to find the best solution of $\boldsymbol{\beta}^\star$ and then update $\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^\star$;

6: Solve (26) using $\mathbf{w}^{(i)}$, $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\beta}^{(i)}$, $\boldsymbol{\Psi}_r^{(i)}$, and $\boldsymbol{\Psi}_t^{(i)}$ to find the solution $(\mathbf{f}^\star, \mathbf{b}^\star, \mathbf{p}^\star)$ and then update $(\mathbf{f}^{(i+1)}, \mathbf{b}^{(i+1)}, \mathbf{p}^{(i+1)}) = (\mathbf{f}^\star, \mathbf{b}^\star, \mathbf{p}^\star)$;

7: Solve (27) using $\mathbf{w}^{(i)}$, $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\beta}^{(i)}$,$\mathbf{b}^{(i)}$,$\mathbf{p}^{(i)}$, and $\mathbf{f}^{(i)}$ to find the best solution of $(\boldsymbol{\Psi}_r^\star, \boldsymbol{\Psi}_t^\star)$ and then update $(\boldsymbol{\Psi}_r^{(i+1)}, \boldsymbol{\Psi}_t^{(i+1)}) = (\boldsymbol{\Psi}_r^\star, \boldsymbol{\Psi}_t^\star)$;

8: Set $i := i + 1$;

9: **until** Convergence or $i > I^{max}$.

aforementioned sub-problems. The solution output from each algorithm in the current iteration is used as input for the other algorithm in an alternating manner. This process is repeated until either a stationary point is reached or the maximum number of iterations I_{max} is reached.

F. Verification of AO convergence:

The sequence $\{\mathcal{X}^{(j)}\}$ generated by the AO algorithm, with

$$\mathcal{X}^{(j)} = (\mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\beta}^{(j)}, \mathbf{b}^{(j)}, \mathbf{p}^{(j)}, \boldsymbol{\Phi}^{(j)}, f^{\text{un }(j)}, f^{\text{ecs}(j)}),$$

converges to a stationary point of the optimization problem, assuming that each subproblem is convex and the objective function $f(\mathcal{X})$ is continuously differentiable within a compact feasible set. 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 [50] 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 [51], 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.

Algorithm 2 SGD-based Algorithm for Solving (15)

1: **Initialize:** Set iteration $i = 0$, maximum iterations $I^{\max} = 1000$, learning rate η.

2: Randomly initialize parameters $\mathbf{w}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\mathbf{b}^{(0)}$, $\mathbf{p}^{(0)}$, $\mathbf{f}^{(0)}$, $\boldsymbol{\Psi}_r^{(0)}$, $\boldsymbol{\Psi}_t^{(0)}$.

3: **repeat**

4: Calculate the optimization function $L = \sum_{j=1}^{J} T_j^{\text{tot}}$.

5: Compute the gradient of L w.r.t each parameter:

$$\nabla_{\mathbf{w}} L = \frac{\partial L}{\partial \mathbf{w}}, \nabla_{\boldsymbol{\mu}} L = \frac{\partial L}{\partial \boldsymbol{\mu}}, \nabla_{\boldsymbol{\beta}} L = \frac{\partial L}{\partial \boldsymbol{\beta}}, \nabla_{\mathbf{b}} L = \frac{\partial L}{\partial \mathbf{b}},$$
$$\nabla_{\mathbf{p}} L = \frac{\partial L}{\partial \mathbf{p}}, \nabla_{\mathbf{f}} L = \frac{\partial L}{\partial \mathbf{f}}, \nabla_{\boldsymbol{\Psi}_r} L = \frac{\partial L}{\partial \boldsymbol{\Psi}_r}, \nabla_{\boldsymbol{\Psi}_t} L = \frac{\partial L}{\partial \boldsymbol{\Psi}_t}.$$

6: Update each parameter using the computed gradient:

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - \eta \nabla_{\mathbf{w}} L, \boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^{(i)} - \eta \nabla_{\boldsymbol{\mu}} L,$$
$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \eta \nabla_{\boldsymbol{\beta}} L, \mathbf{b}^{(i+1)} = \mathbf{b}^{(i)} - \eta \nabla_{\mathbf{b}} L,$$
$$\mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} - \eta \nabla_{\mathbf{p}} L, \mathbf{f}^{(i+1)} = \mathbf{f}^{(i)} - \eta \nabla_{\mathbf{f}} L,$$
$$\boldsymbol{\Psi}_r^{(i+1)} = \boldsymbol{\Psi}_r^{(i)} - \eta \nabla_{\boldsymbol{\Psi}_r} L, \boldsymbol{\Psi}_t^{(i+1)} = \boldsymbol{\Psi}_t^{(i)} - \eta \nabla_{\boldsymbol{\Psi}_t} L.$$

7: $i \leftarrow i + 1$

8: **until** $i > I^{\max}$ or convergence

G. Benchmark Solution

We have also integrated a stochastic gradient descent (SGD) algorithm as a benchmark solution specifically designed for online implementation. The SGD approach is well-suited for environments requiring continuous and real-time adaptation, making it highly appropriate for the dynamic nature of IoT networks. This method not only accommodates the limited computational resources of small cells and IoT devices but also maintains the solution's practicality in live scenarios.

Algorithm 2 describes an SGD-based approach for optimizing system parameters by iteratively calculating gradients of an optimization function and updating parameters using a specified learning rate. This process is repeated until either the maximum number of iterations is reached or convergence is achieved, aiming to optimize the total e2e latency across tasks in the system.

H. Computational Complexity Analysis

We have conducted a detailed computational complexity analysis for both AO and SGD. This analysis provides a critical comparison of the computational demands and efficiency of each optimization technique, thereby illuminating their practical applicability in various scenarios. The more details of the computational complexity analysis are provided as follows:

TABLE III: Computational complexities of subproblems in AO.

Variable	Computational Complexity
$\mathbf{w}$	$O\left(J^2\sqrt{4J}\right)$
$\boldsymbol{\mu}$	$O\left(J^2\sqrt{2J+2}\right)$
$\boldsymbol{\beta}$	$O\left(J^2\sqrt{3J+1}\right)$
$\mathbf{b}, \mathbf{p}, \mathbf{f}$	$O\left(16J^2\sqrt{7J+2}\right)$
$\boldsymbol{\Psi}_r, \boldsymbol{\Psi}_t$	$O\left(4N^2\sqrt{2N+1}\right)$

TABLE IV: Simulation parameters.

Parameter	Value	Parameter	Value	Parameter	Value
M, N	4, 64	$F_{\max}^{\text{ecs}}$	30 GHz	$R_{\min}$	0.3 bit/s
J_r, J_t	15, 15	$T_j^{\max}$	10 ms	$E_j^{\max}$	3 mJ
P_{max}^{RIS}	9 dBW	B	5 MHz	ε_j	10^{-8}

1) Computational Complexity analysis of AO algorithm: The overall computational complexity for the alternating optimization process is attributed to the updates of the variables $\mathbf{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, $\mathbf{b}$, $\mathbf{p}$, $\mathbf{f}$, $\boldsymbol{\Psi}_r$, and $\boldsymbol{\Psi}_t$. The computational complexity of the 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. The overall computational complexity of the AO algorithm can be approximated as $O(J^2\sqrt{4J}+J^2\sqrt{2J+2}+J^2\sqrt{3J+1}+16J^2\sqrt{7J+2}+4N^2\sqrt{2N+1})$.

2) Computational Complexity Analysis of the SGD Algorithm: The computational complexity of the SGD algorithm, as outlined in Algorithm 2, is analyzed based on its iterative process, where each iteration involves computing the gradient of a loss function and updating several parameters. Each iteration begins with the computation of gradients for eight parameters: $\mathbf{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, $\mathbf{b}$, $\mathbf{p}$, $\mathbf{f}$, $\boldsymbol{\Psi}_r$, and $\boldsymbol{\Psi}_t$. Assuming the average dimensionality of each parameter is d. Each partial derivative calculation has a complexity of $O(d)$, the gradient computation for all parameters collectively within a single iteration sums up to $O(\sum_{v=1\in\mathcal{V}}\eta\cdot d_v\cdot d_c)$, where v represents the number of terms in the loss function and η is the learning rate. Following gradient computation, parameter updates involve simple arithmetic operations i.e., multiplication by a learning rate and subtraction, each with a complexity of $O(d_v)$ per parameter, thus $O(\sum_{v=1\in\mathcal{V}}d_v)$ for all parameters. The overall per-iteration complexity is dominated by the gradient computation, hence estimated at $O(\eta\cdot(J(4J)+J(2J+2)+J(3J+1)+4J(7J+2)+2N(2N+1)))$.

IV. 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 [8], [49] is utilized, as listed in **Table IV**.

Fig. 3 illustrates the convergence of the active STAR-RIS-assisted and FD relay-assisted DT-based URLLC system with varying values of N under perfect and imperfect CSI conditions. Both systems show a sharp decline in latency during the initial iterations, attributed to the rapid learning phase as they adapt to propagation conditions. The initial steep gradient indicates significant benefits from adjustments in reflective elements or relay amplification. After this drop, some configurations plateau, suggesting equilibrium where further iterations yield minimal latency improvements, likely due to optimization algorithm limits and system constraints. Non-monotonic behavior is observed, with small rises in latency at certain iterations reflecting the complexity of the optimization process, where specific adjustments temporarily degrade performance before converging optimally. Increased latency under imperfect CSI highlights challenges from less

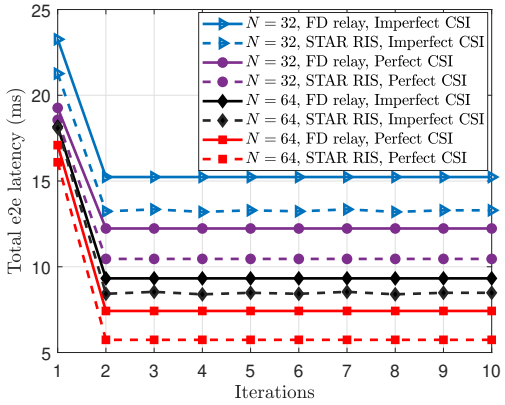

Fig. 3: Convergence of benchmark schemes.

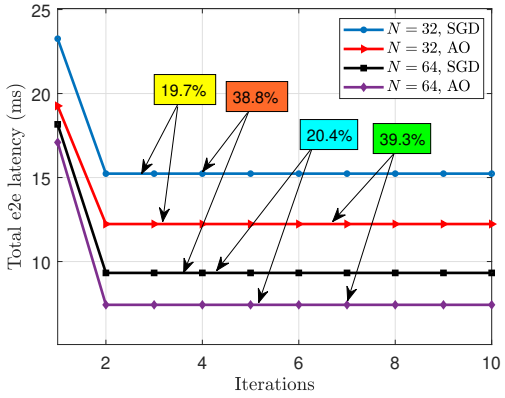

Fig. 4: Convergence of AO and SGD.

accurate channel information, affecting decision-making and preventing optimal latency. Results show that increasing N (from 32 to 64) decreases latency in both scenarios, attributed to enhanced signal control from more reflective elements, which effectively shape signals, mitigate interference, and optimize quality, leading to minimized delays and lower latency.

Fig 4 compares the convergence of the AO and SGD algorithms for varying RIS elements (N). The AO algorithm consistently outperforms SGD, achieving latency reductions of 19.7% for $N = 32$ and 20.4% for $N = 64$. This performance is due to AO's ability to leverage problem structure and optimize parameters iteratively, leading to more effective convergence. Unlike SGD, which updates parameters based on random samples, AO refines solutions by sequentially optimizing components, thereby better exploiting the problem's inherent structure and avoiding suboptimal convergence points. The superiority of AO is highlighted by its latency reduction of 39.3% when transitioning from $N = 32$ to $N = 64$, compared to 38.8% for SGD. This difference, while marginal, is significant in large-scale systems, where even small improvements can enhance responsiveness and efficiency. AO's adaptability to system configurations allows for finer adjustments and precise control, especially as N increases and optimization complexity rises, ensuring consistently lower latency.

Fig. 5 illustrates the convergence of three operational modes in the active STAR-RIS system: energy splitting (ES), time

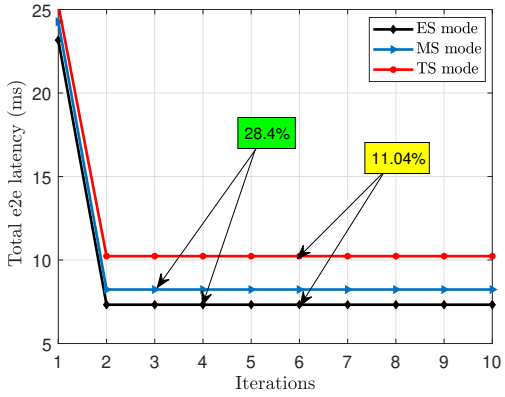

Fig. 5: Impact of different working modes.

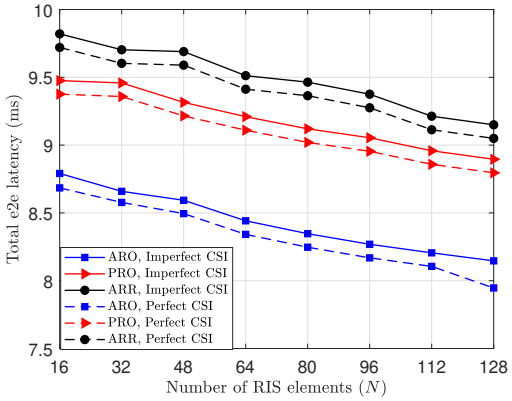

Fig. 7: Impact of number of RIS elements (N).

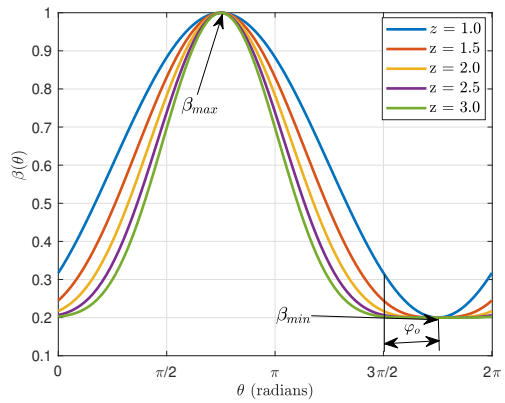

Fig. 6: Reflection amplitude variations with phase shift in active STAR-RIS.

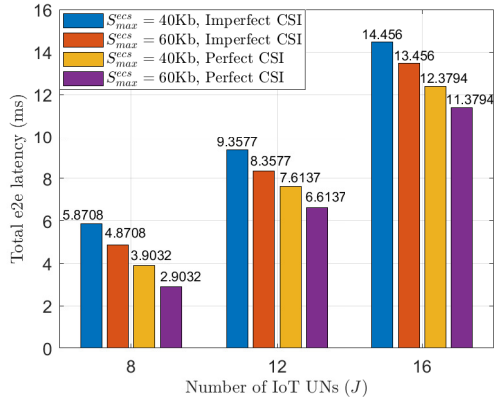

Fig. 8: Impact of S^{ecs}_{max}.

switching (TS), and mode switching (MS). The results indicate that the ES mode significantly reduces latency, achieving a 28.4% reduction compared to the MS mode and an 11.04% reduction against the TS mode. This advantage stems from the ES mode's dual-functional capability, allowing simultaneous energy harvesting and data transmission, thereby optimizing signal management. In the ES mode, the incoming signal is divided into two pathways—one for energy harvesting and the other for information transmission—maximizing electromagnetic wave utilization without switching. In contrast, the TS mode alternates between functions, leading to inactivity and latency during transitions. The MS mode, which dynamically switches based on system demands, adds complexity and potential timing errors, further worsening latency issues. Thus, the ES mode's seamless integration of functions simplifies the system architecture while enhancing efficiency and responsiveness.

Fig. 6 illustrates the reflection amplitude variations $(\beta_n(\varphi_{p,n}))$ as a function of the phase shift $(\varphi_{p,n})$ in an active STAR-RIS, plotted over 0 to 2π radians. To establish the crucial relationship between reflection amplitude and phase shift for designing active STAR-RIS-aided wireless systems, we utilize an analytical model applicable to various semiconductor devices used in STAR-RIS [52]. We define $v_{p,n} = \beta_n(\varphi_{p,n})\, e^{j\varphi_{p,n}}, \forall p \in \{r, t\}$, where $\varphi_{p,n} \in [0, 2\pi)$ and $\beta_n(\varphi_{p,n}) \in [0, 1]$ denote the phase shift and corresponding amplitude for each STAR-RIS element. The relationship for $\beta_n(\varphi_{p,n})$ is expressed as:

$$\beta_n(\varphi_{p,n}) = \beta_{\min} + (1 - \beta_{\min}) \left(\frac{\sin(\varphi_{p,n} - \varphi_o) + 1}{2} \right)^z,$$

where $\beta_{\min} \geq 0$ denotes the minimum amplitude, and $z \geq 0$ controls the steepness of the response curve. We select $\beta_{\min} = 0.2$ and $\beta_{\max} = 1$ to reflect realistic operational limits, while the phase offset $\varphi_o = \frac{\pi}{4}$ aligns with typical active element responses. This model reflects practical constraints and operational characteristics of active STAR-RIS elements, which adjust phase and amplitude based on real-time conditions. The parameters $\beta_{\min}$, ϕ, and z can be determined through standard curve fitting, facilitating precise optimization of the STAR-RIS for improved communication reliability and efficiency. The figure emphasizes the importance of phase-dependent amplitude variations in active STAR-RIS design, justifying adaptable beamforming strategies essential for optimizing operational efficacy across varying conditions.

Fig. 7 illustrates a comparative analysis of the proposed system under perfect and imperfect CSI scenarios, utilizing active STAR-RIS optimal beamforming (ARO) across varying N. This is compared with benchmark schemes: passive STAR-RIS with optimal beamforming (PRO) and active STAR-RIS with random beamforming (ARR). The ARO achieves a ≈10%

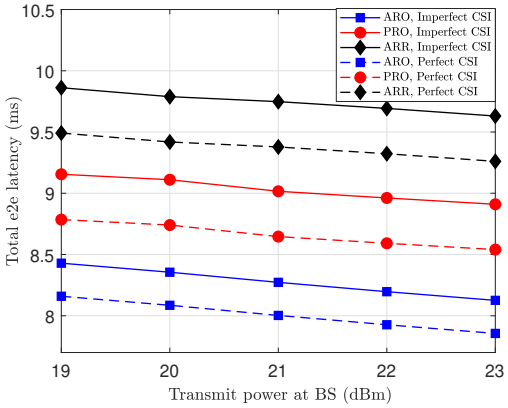

Fig. 9: Impact of transmit power at BS (p).

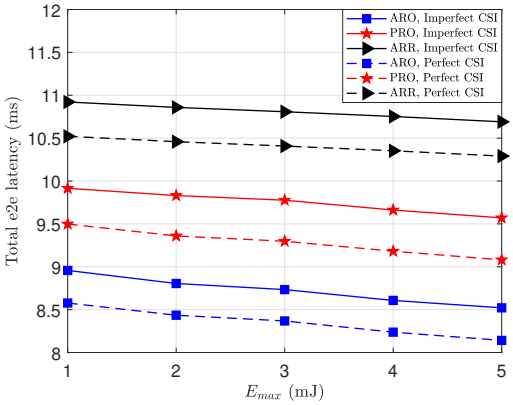

Fig. 10: Impact of E_{max}.

latency reduction over the PRO due to its ability to adjust phase shifts, optimize beamforming, and mitigate interference. The PRO also outperforms the ARR, achieving a ≈9% latency reduction through its optimized beamforming strategy, which leverages prior channel knowledge. The number of active STAR-RIS elements directly impacts latency, as their integration enhances spectrum capacity, reduces interference, and improves overall system performance. High CSI accuracy allows IoT UNs to adapt to channel fluctuations, minimizing data transmission times, thus enabling lower latency in perfect CSI scenarios through effective communication channel management.

Fig. 8 demonstrates the total e2e latency versus J by varying S_{max}^{ecs} under perfect and imperfect CSI cases. The increase in S_{max}^{ecs} from 40 Kb to 60 Kb for $J = 8, J = 12$, and $J = 16$ results in a respective reduction of total e2e latency by 22.24 %, 16.39 %, and 16.97 % under perfect CSI scenario, and by 50.9 %, 22.9 %, and 16.77 % under imperfect CSI scenario. Firstly, the trend of latency reduction with increased S_{max}^{ecs} occurs because higher edge caching capacity allows more efficient data storage and retrieval in the ECS, thereby reducing the need for data transmission from IoT UN to the ECS. Consequently, the decreased data transmission requirements lead to lower latency, improving system performance. Secondly, as J increases, latency also increases. This is due to the higher competition for network resources with an increasing number of IoT UNs, leading to higher congestion and potential delays in data transmission. Thirdly, perfect CSI yields lower latency results because high CSI accuracy allows IoT UNs to respond precisely to channel fluctuations, providing optimal adaptive capabilities and minimizing the time required to transmit and receive data.

Fig. 9 illustrates the effect of $p = p_1 = p_2 = \cdots = p_J$ on the total end-to-end latency of the proposed system. Increasing power levels enhances signal transmissions from IoT UNs to the BS, improving signal quality and reducing interference. Although higher transmit power may raise interference, the system employs advanced techniques like scheduling, power control, and beamforming to mitigate this. Under equal power budgets, PRO with perfect CSI shows a ≈ 7.41% latency reduction compared to ARR, while ARO with perfect CSI achieves a ≈ 42.56% lower latency than PRO. In scenarios with imperfect CSI, PRO still delivers a ≈ 4.76% lower latency than ARR, and ARO demonstrates a ≈ 23.08% lower latency than PRO. This improved signal quality and reduced interference lead to fewer transmission issues and a significant decrease in total latency. Utilizing perfect CSI allows for precise adjustments based on accurate channel information, reducing delays and enhancing the efficiency of wireless communication.

Fig. 10 depicts the total e2e latency versus variations in E_{max} under the conditions of perfect CSI and imperfect CSI. As E_{max} ranges from 1 to 5, ARO demonstrates a ≈ 6.54% and ≈ 55.07% reduction in total e2e latency under perfect CSI conditions compared to PRO and ARR, respectively, while experiencing corresponding reductions of ≈ 5.66% and ≈ 51.94% under imperfect CSI conditions compared to PRO and ARR. This phenomenon is attributed to the intricate trade-off between energy efficiency and responsive performance, highlighting the need for a balanced approach in managing the relationship between energy and latency. Increased energy can enhance processing speed and throughput, potentially reducing latency by allowing the system to respond more quickly to demands. On the other hand, the role of perfect CSI and imperfect CSI also influences latency. Perfect CSI demonstrates lower latency results in contrast to imperfect CSI conditions.

Fig. 11 highlights the effect of the minimum rate on the total e2e latency under both perfect CSI and imperfect CSI scenarios, considering the benchmark schemes, i.e., PRO and ARR. Increasing the minimum rate from 0.3 bit/s to 0.7 bit/s reduces latency. Under perfect CSI conditions, ARO achieves a 9.8% and 22.6% reduction in total e2e latency compared to PRO and ARR, respectively, while under imperfect CSI, the corresponding reductions are 10.4% and 22.7% when compared to PRO and ARR. The observed reduction in latency by ARO is attributed to the influence of a higher data transfer rate, which facilitates faster movement of information between the IoT UNs and the BS. The increased transfer rate enables quicker data transmission from one point to another within the network. As a result, the overall time required for data to transfer from the IoT UNs to the BS is reduced, leading to a

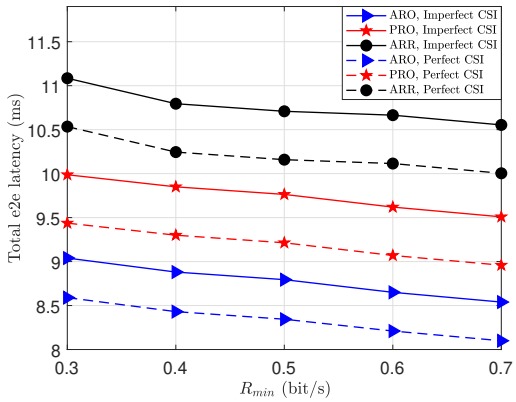

Fig. 11: Impact of minimum rate (R_{min}).

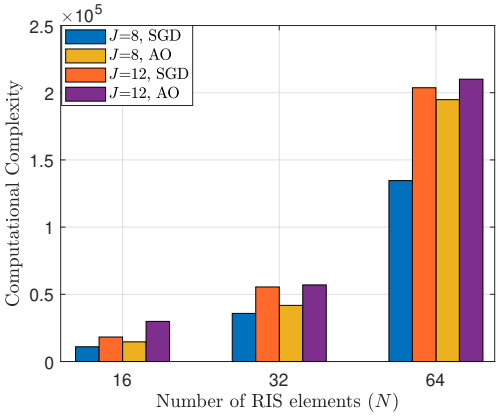

Fig. 13: Computational complexity versus number of RIS elements.

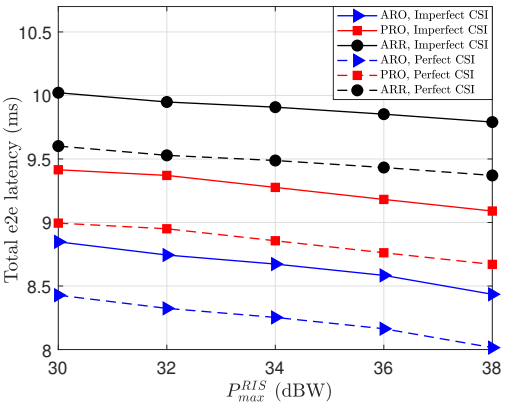

Fig. 12: Impact of maximum power at RIS.

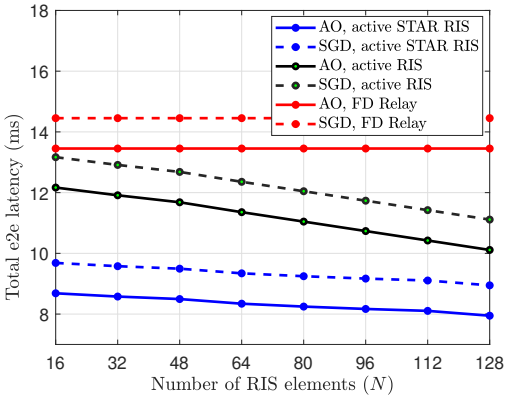

Fig. 14: Total e2e latency performance comparison of active STAR-RIS, active RIS, and FD relay using AO and SGD algorithms.

notable decrease in latency. This underscores the crucial role of enhanced data transfer rates in optimizing the efficiency of communication systems, particularly in the context of IoT networks. In contrast to imperfect CSI, perfect CSI can lower latency by enabling the transmitter to make precise adjustments based on accurate knowledge of communication channel conditions. This reduces the risk of transmission errors, speeds up data delivery between the transmitter and receiver, and allows the transmitter to optimize transmission strategies, minimizing delays and improving overall efficiency in wireless communication.

Fig. 12 highlights the impact of increasing power at active STAR-RIS on total e2e latency under perfect CSI and imperfect CSI scenarios. The augmentation of RIS power enhances the strength of the transmitted signal, leading to a more reliable transmission between the IoT UNs and the BS. This reduction in the time required to transmit data and overall improved system responsiveness contribute to reduced latency. The findings reveal notable performance improvements in the perfect CSI case compared to the imperfect CSI case. Specifically, there is a 4.7% decrease in ARO performance under a perfect CSI scenario compared to an imperfect CSI scenario, underscoring the positive impact on reliability. Both PRO and ARR also experience a 4.4% reduction each, highlighting enhanced efficiency in packet reception and round-trip communication. These results emphasize the significance of obtaining accurate CSI in optimizing various performance metrics, showcasing the potential for refined communication systems and network reliability.

Fig. 13 illustrates a comparative analysis of the computational complexities of the SGD and AO algorithms, evaluated at different numbers of active STAR-RIS elements (denoted by $N = 16, 32, 64$) and number of IoT UNs ($J = 8, 12$). It is evident from the figure that the SGD algorithm consistently exhibits lower computational complexity compared to the AO algorithm across all values of N. This trend is attributed to the inherent algorithmic designs: SGD's complexity formula scales linearly with J and polynomially with N, while AO's complexity involves quadratic terms in both J and N, compounded by square roots of polynomial expressions, leading to a steeper increase in complexity. Moreover, it is observed that as the number of IoT UNs increases, the computational burden for both algorithms also rises, which is indicative of the scalability challenges in systems with a large user base. The lower complexity of SGD suggests its suitability for scenarios requiring rapid processing and limited computational resources. In contrast, the higher complexity of AO may offer better optimization performance, which is suitable for applications where solution quality is more critical than computational speed.

Fig. 14 compares the end-to-end (e2e) latency performance

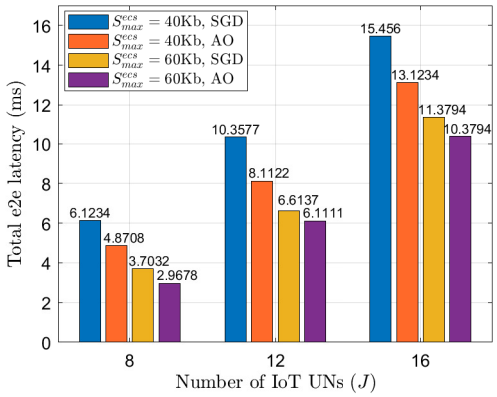

Fig. 15: Impact of edge caching capacity on total e2e latency for AO and SGD with varying numbers of IoT UNs

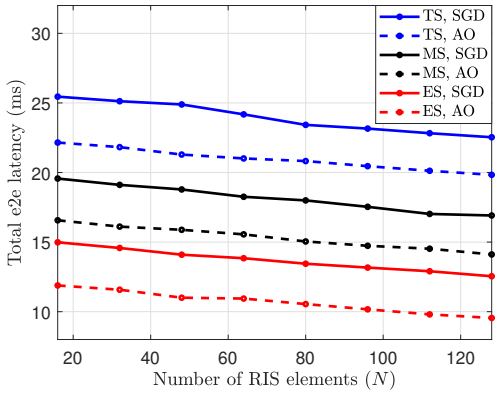

Fig. 16: Impact of different working modes versus number of RIS elements for AO and SGD algorithms.

of active STAR-RIS, active RIS, and FD relay-assisted DT-MEC-URLLC systems using AO and SGD algorithms. The results show that the active STAR-RIS system significantly outperforms the others, achieving lower latency than both active RIS and FD relay systems. This advantage arises from active STAR-RIS's advanced signal management, which optimizes the propagation environment by modifying both amplitude and phase of incoming signals in real-time over a 360-degree coverage. The AO algorithm further enhances active STAR-RIS's physical capabilities by iteratively fine-tuning parameters based on real-time feedback, while traditional active RIS systems, despite their ability to manipulate signal phases, lack simultaneous amplitude adjustments, limiting their control over the signal path. Although FD relays extend coverage and improve signal strength, they do not match the dynamic environmental adaptation of STAR-RIS systems, resulting in higher latency. Notably, the latency of FD relay systems (both AO and SGD) remains constant with increasing N, indicating that their performance relies on direct transmission methods rather than the reflective and refractive properties of RIS.

Fig 15 illustrates the total end-to-end latency versus J by varying the S^{ecs}_{max} for AO and SGD. Increasing the cache size from 40kB to 60kB results in lower latency for both SGD and AO at each level of J. This outcome suggests that the larger edge caching size of 60 Kb contributes to lower latency due to its ability to store and serve more content closer to the end-users. With a larger cache capacity, there is a higher probability of caching popular or frequently requested content, reducing the need to fetch content from distant servers. As a result, the overall latency is decreased, improving user experience and network efficiency. The larger cache size likely involves managing more data, which could introduce delays in data retrieval and processing. The increase in latency with more IoT UNs is primarily due to higher data volume, which strains network resources, and increased network traffic, leading to congestion and delays. Additionally, more UNs cause resource contention at network and edge servers, further slowing down data processing and increasing latency. These factors collectively escalate the complexity of network management, exacerbating latency challenges in dense IoT environments.AO consistently demonstrates lower latencies compared to SGD across all scenarios. The reduction in latency suggests that AO is more efficient in optimizing the system's parameters under varying load and cache conditions.

Fig. 16 demonstrates the impact of the number of RIS elements on the latency for different operational modes i.e., ES, MS, and TS—utilizing AO and SGD algorithms. As the number of RIS elements increases, there is a notable reduction in latency across all modes and algorithms, suggesting that a higher count of RIS elements facilitates improved performance in active STAR-RIS-enhanced communication systems. A key observation is the performance disparity between the AO and SGD algorithms. While the AO algorithm shows superior latency performance, particularly in the ES mode, the SGD algorithm, despite being effective, trails in efficiency, as indicated by the higher latency lines for both MS and TS modes. This can be attributed to the inherent nature of SGD, where optimization might not be as dynamically responsive to changes in the system's state as AO. In the MS mode, while AO and SGD both manage to decrease latency as the number of RIS elements increases, the decrease is more gradual compared to the ES mode. This points to the potential overhead and complexity introduced by mode switching, which might not be as latency-efficient as continuous energy and data management seen in the ES mode. The TS mode, utilizing time slots for energy harvesting and data transmission, shows the highest latency among the three modes for both algorithms. This could be due to the operational inefficiencies associated with the switching process, which can introduce delays and reduce the system's overall responsiveness and speed.

V. Conclusions

This paper has investigated an active STAR-RIS-assisted DT-based MEC system for the first time, facilitating task offloading and enhancing both IoT-URLLC services and spatial coverage. To this end, we formulated a comprehensive optimization problem to minimise total e2e latency while considering active STAR-RIS and MEC constraints. We then solved this non-convex optimization problem using an efficient AO algorithm and compared the outcomes with the SGD

algorithm. The results depict that AO algorithm consistently delivers superior results compared to SGD, achieving latency reductions of 19.7% at $N = 32$ and 20.4% at $N = 64$. Moreover, enhancing N from 32 to 64 yields a significant latency reduction of 39.3% with AO, slightly better than SGD's 38.8%. While AO provides greater latency reductions, it is important to note that SGD maintains a consistently lower computational complexity throughout. Additionally, the ES mode further reduces the system's total e2e latency by 28.4% compared to the MS mode and 11.04% compared to the TS mode. Our results also demonstrated noteworthy improvements in total e2e latency when varying edge caching scenarios (S_{max}^{ecs}). Under perfect CSI, increasing S_{max}^{ecs} led to total e2e latency reductions of $22.24\%, 16.39\%$, and 16.97% for $J = 8$, $J = 12$, and $J = 16$, while under imperfect CSI, reductions were $50.9\%, 22.9\%$, and 16.77%. Elevating E_{max} from 1 to 5 resulted in ARO achieving $\approx 6.54\%$ and $\approx 55.07\%$ latency reduction compared to PRO and ARR under perfect CSI, and $\approx 5.66\%$ and $\approx 51.94\%$ reduction under imperfect CSI. Under perfect CSI, ARO vs. PRO and ARO vs. ARR yielded 9.86% and 22.6% latency reduction, while under imperfect CSI, reductions were 10.47% and 22.6%. Moreover, assessing RIS maximum power impact, ARO, PRO, and ARR showed $4.7\%, 4.4\%$, and 4.4% latency decrease in perfect CSI scenario versus imperfect CSI, highlighting active STAR-RIS's superior performance over passive STAR-RIS.

REFERENCES

[1] 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.

[2] S. Kurma, P. K. Sharma, K. Singh, S. Mumtaz, and C.-P. Li, "URLLC-based cooperative industrial IoT networks with nonlinear energy harvesting," *IEEE Trans. Industr. Inform.*, vol. 19, no. 2, pp. 2078–2088, Feb. 2023.

[3] S. Kurma, P. K. Sharma, S. Dhok, K. Singh, and C.-P. Li, "Adaptive AF/DF two-way relaying in FD multi-user URLLC system with user mobility," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10 224–10 241, Dec. 2022.

[4] 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.

[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] 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.

[8] 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.

[9] 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.

[10] S. Kurma, M. Katwe, K. Singh, C. Pan, 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)*, Hoboken, NJ, USA, May 2023, pp. 1–6.

[11] ——, "RIS-empowered MEC for URLLC systems with digital-twin-driven architecture," *IEEE Trans. Commun.*, pp. 1–1, May 2023.

[12] S. Kurma, M. Katwe, K. Singh, T. Q. Duong, and C.-P. Li, "Spectral-energy efficient resource allocation in RIS-aided FD-MIMO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Oct. 2023.

[13] S. Kurma, K. Singh, P. K. Sharma, and C.-P. Li, "DRL approach for spectral-energy trade-off in RIS-assisted full-duplex multi-user MIMO systems," in *Proc. IEEE Wireless Commun. Netw. (WCNC),*, May 2023, pp. 1–6.

[14] 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.

[15] S. Zhang, W. Hao, G. Sun, C. Huang, Z. Zhu, X. Li, and C. Yuen, "Joint beamforming optimization for active STAR-RIS assisted ISAC systems," *arXiv preprint arXiv:2308.06064*, 2023.

[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 *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, 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] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] C. Wu, Y. Liu, X. Mu, X. Gu, and O. A. Dobre, "Coverage characterization of star-ris networks: Noma and oma," *IEEE Commun.Lett.*, vol. 25, no. 9, pp. 3036–3040, Sep. 2021.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] S. Kurma, T. A. Lestari, K. Singh, A. Paul, and S. Mumtaz, "Active RIS in digital twin-based URLLC IoT networks: Fully-connected vs. sub-connected?" *IEEE Trans. Wireless Commun.*, pp. 1–1, Apr. 2024.

[38] D. Van Huynh, V.-D. Nguyen, V. Sharma, O. A. Dobre, and T. Q. Duong, "Digital twin empowered ultra-reliable and low-latency communications-based edge networks in industrial iot environment," in *Prof. ICC IEEE International Conference on Communications*. IEEE, 2022, pp. 5651–5656.

[39] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing offloading latency for digital twin edge networks in 6g," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 240–12 251, 2020.

[40] T. Liu, L. Tang, W. Wang, Q. Chen, and X. Zeng, "Digital-twin-assisted task offloading based on edge collaboration in the digital twin edge network," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1427–1444, Jan. 2022.

[41] 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.

[42] Z. Zhou, Z. Jia, H. Liao, W. Lu, S. Mumtaz, M. Guizani, and M. Tariq, "Secure and latency-aware digital twin assisted resource scheduling for 5G edge computing-empowered distribution grids," *IEEE Trans. Industr. Inform.*, vol. 18, no. 7, pp. 4933–4943, July 2022.

[43] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 2, pp. 1035–1071, Feb. 2022.

[44] A. Papazafeiropoulos, H. Q. Ngo, P. Kourtessis, and S. Chatzinotas, "STAR-RIS assisted cell-free massive MIMO system under spatially-correlated channels," *arXiv:2311.18343*, 2023.

[45] A. Papazafeiropoulos, H. Ge, P. Kourtessis, T. Ratnarajah, S. Chatzinotas, and S. Papavassiliou, "Two-timescale design for active STAR-RIS aided massive MIMO systems," *IEEE Trans. Veh. Technol.*, pp. 1–16, 2024.

[46] N. Garg, A. K. Jagannatham, G. Sharma, and T. Ratnarajah, "Precoder feedback schemes for robust interference alignment with bounded CSI uncertainty," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 407–425, 2020.

[47] H. Zhang and L. Hanzo, "Federated learning assisted multi-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14 104–14 109, Nov. 2020.

[48] 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.

[49] 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.

[50] 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.

[51] C. R. Alcantud, Jose, "Softarisons: theory and practice," *Neural Computing and Applications*, vol. 33, no. 23, pp. 16 759–16 771, 2021.

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

Tri Ayu Lestari (Graduate Student Member, IEEE) She earned her Associate degree in Telecommunication Engineering with Cum Laude honors from the State Polytechnic of Padang, Indonesia, in 2019, followed by a Bachelor of Applied Science in Telecommunication Engineering with Cum Laude from the Electronic Engineering Polytechnic Institute of Surabaya (EEPIS), Indonesia, in 2022. In 2024, she completed her Master's degree in the International Master Program in Telecommunication Engineering (IMPTE) at National Sun Yat-Sen University (NSYSU), Taiwan. Her research focuses on reconfigurable intelligent surfaces, Digital Twin, and ultra-reliable low latency communication.

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. In 2024, She received Ph.D in the Institute of Communications Engineering (ICE) from 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, large language models (LLMs) and machine learning for communication. She received the Excellent Ph.D Student in Research Award in 2024.

Anal Paul (Member, IEEE) received his Bachelor of Technology degree from the Government College of Engineering and Ceramic Technology, India, in 2008, and his Master of Engineering degree from Jadavpur University, India, in 2010. In 2021, he received his Ph.D. degree from the Indian Institute of Engineering Science and Technology, Shibpur. From July to December 2022, he worked as a postdoctoral researcher in the Department of Information and Communication Engineering at Yeungnam University, South Korea. Since January 2023, he has been a Postdoctoral Researcher at National Sun Yat-sen University, Taiwan, conducting research in Digital Twin and Metaverse applications for Wireless Communication Systems.

Keshav Singh (Member, IEEE) received the Ph.D. degree in Communication Engineering from National Central University, Taiwan, in 2015. He is currently with the Institute of Communications Engineering, National Sun Yat-sen University (NSYSU), Taiwan, as an Associate Professor. He is also an Adjunct Professor at the Memorial University, Canada. 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 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.

Dr. Singh chaired workshops on conferences like IEEE GLOBECOM 2023 and IEEE WCNC 2024. He also serves as leading guest editor for IEEE Transactions on Green Communications and Networking Special Issue on Design of Green Near-Field Wireless Communication Networks and IEEE Internet of Things Journal Special Issue on Positioning and Sensing for Near-Filed (NF)-driven Internet-of-Everything.

Simon L. Cotton (S'04–M'07–SM'14) received the B.Eng. degree in electronics and software from Ulster University, Ulster, U.K., in 2004, and the Ph.D. degree in electrical and electronic engineering from the Queen's University of Belfast, Belfast, U.K., in 2007. From 2007 to 2011 he was a Research Fellow, then Senior Research Fellow, 2011 to 2012, Lecturer (Assistant Professor), 2012 to 2015, and Reader (Associate Professor), 2015 to 2019 at the Queen's University of Belfast. He is currently a Full Professor and the Director of the Centre for Wireless Innovation (CWI), Queen's University Belfast. Professor Cotton has authored and co-authored over 180 publications in major IEEE/IET journals and refereed international conferences, three book chapters, and two patents. Among his research interests are propagation measurements and statistical channel characterization. His other research interests include cellular device-to-device, vehicular, and body-centric communications. Professor Cotton was awarded the H. A. Wheeler Prize, in July 2010, by the IEEE Antennas and Propagation Society for the best applications journal paper in the IEEE Transactions on Antennas and Propagation during 2009. In July 2011, he was awarded the Sir George Macfarlane Award from the U.K. Royal Academy of Engineering in recognition of his technical and scientific attainment since graduating from his first degree in engineering.

Trung Q. Duong (Fellow, IEEE) is a Canada Excellence Research Chair (CERC) and a Full Professor at Memorial University, Canada. He is also the adjunct Chair Professor in Telecommunications at Queen's University Belfast, UK and a Research Chair of Royal Academy of Engineering, UK. He was a Distinguished Advisory Professor at Inje University, South Korea (2017-2019), an Adjunct Professor and the Director of Institute for AI and Big Data at Duy Tan University, Vietnam (2012-present), and a Visiting Professor (under Eminent Scholar program) at Kyung Hee University, South Korea (2023-2025). His current research interests include quantum communications, wireless communications, quantum machine learning, and quantum optimisation.

Dr. Duong has served as an Editor/Guest Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS MAGAZINES, and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He received the Best Paper Award at the IEEE VTC-Spring 2013, IEEE ICC 2014, IEEE GLOBECOM 2016, 2019, 2022, IEEE DSP 2017, IWCMC 2019, 2023, and IEEE CAMAD 2023. He has received the two prestigious awards from the Royal Academy of Engineering (RAEng): RAEng Research Chair (2021-2025) and the RAEng Research Fellow (2015-2020). He is the recipient of the prestigious Newton Prize 2017.