Joint Offloading Policy and Resource Allocation in IRS-aided MEC for IoT Users with Short Packet Transmission

Jalal Jalali†, Ata Khalili‡, Rafael Berkvens†, and Jeroen Famaey†
†IDLab, University of Antwerp - imec, Sint-Pietersvliet 7, 2000 Antwerp, Belgium
‡Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg, Erlangen, Germany
Email: jalal.jalali@uantwerpen.be

2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall) | 979-8-3503-2928-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/VTC2023-Fall60731.2023.10333867

***Abstract*—This paper focuses on leveraging a mobile edge computing (MEC) server at an access point (AP) to address the delay and reliability sensitivity requirement of multi-user machine-type communication (MTC). By offloading tasks to the MEC server, latency for low-power MTC devices can be minimized. Meanwhile, intelligent reflecting surfaces (IRSs) are supported to facilitate robust offloading, enhance spectrum efficiency, and improve coverage by influencing incident radio-frequency wave propagation via modifying the phase shifts with passive reflecting components. Therefore, we investigate joint radio resource allocation and edge offloading decision optimization in a multi-user IRS-assisted MEC network, wherein a multi-antenna AP receives information symbols from a set of Internet of Things (IoT) users with short packet transmission. In particular, we minimize the system's power utilization subject to offloading MTC-enabled IoT users' quality of service (QoS) requirements, transmit power feasibility, capacity limitation, and IRS phase shift. The non-convex nature of the formulated problem poses a challenge to solving it effectively. To address this, we propose an efficient iterative algorithm based on successive convex approximation (SCA) and a penalty-based approach for handling unit-modulus constraints in the presence of passive reflecting elements at the IRS. Simulation results demonstrate the superior performance of our algorithm compared to other baseline schemes.**

***Index Terms*—Intelligent reflecting surface (IRS), mobile edge computing (MEC), machine-type communication (MTC), short packet transmission.**

I. Introduction

IN RECENT years, the rapid development of Internet of Things (IoT) use cases has led to the emergence of computation-intensive and latency-sensitive applications such as autonomous driving, augmented reality, virtual reality, and unmanned aerial vehicles [1]–[3]. These applications aim to enable real-time interactions between machines and humans or between machines themselves, that is, machine-type communication (MTC). To support these applications, the next generation of wireless networks, i.e., the 6th generation (6G), must accommodate many IoT devices for real-time computation, communication, and control.

Due to cost and size considerations, IoT devices are often constrained by limited battery capacity and low-performance processors. Consequently, a critical challenge in future IoT networks is how to enhance the computational capability of IoT devices to handle intensive computation loads with strict latency requirements [4]. Although cloud computing, with its rich computational resources, has traditionally been the solution, it may introduce significant computational latency due to the remote locations of cloud servers [5]. Therefore, for mission-critical and time-sensitive applications like healthcare, autonomous driving, and the tactical internet, ultra-reliable and low-latency communication (URLLC) has emerged as a promising IoT service category for meeting future network's reliability and latency requirements [6]. In this context, URLLC can achieve decoding error targets lower than 10^{-5} and latency constraints as low as 1 ms. However, the traditional Shannon capacity formula may not suffice within the short packet regime of URLLC-assisted IoT systems [7].

To further tackle the challenges of computational latency in IoT networks, mobile edge computing (MEC) is considered as another promising solution to mitigate network congestion and significantly decrease latency compared to cloud computing [8]. By deploying servers at the network edge, such as cellular base stations (BSs) or WiFi access points (APs), MEC enables the direct offloading of data and computational tasks from IoT devices to the MEC server, effectively enhancing the quality of experience for the end users [9]. In MEC systems, tasks can be classified into different categories based on their dependencies and partitionability, leading to two typical computation-offloading modes: *partial offloading* and *binary offloading*. Binary offloading necessitates complete local execution or offloading of computations, unlike partial offloading, which allows for segmented execution. These modes involve optimizing computation and communication resource allocations to minimize energy consumption [10], computation latency [11], [12], maximize computation rate [13], and improve energy efficiency [14]. However, wireless channel attenuation between APs and devices can degrade task offloading efficiency [15]. To overcome this limitation, the massive multiple-input multiple-output (MIMO) technique has been employed to enhance the efficiency of task offloading in MEC systems [16].

While massive MIMO boosts the offloading efficiency of MEC systems, it is hindered by high energy and hardware costs. To overcome these, the intelligent reflecting surface (IRS) is a cost-effective solution proposed to enhance spectral and energy efficiency in next-generation mobile networks [17]. An IRS is a digitally-controlled meta-surface, consisting of an IRS controller and numerous affordable passive reflecting elements. It does not need any radio-frequency (RF) chains. By

This work was supported by the CHIST-ERA grant SAMBAS (CHIST-ERA-20-SICT-003), with FWO, ANR, NKFIH, and UKRI funding.

smartly tweaking the phase shifts of each IRS element through the IRS controller, it can dynamically alter the wireless propagation environment to achieve various design goals, like signal enhancement or interference reduction [7], [18], [19]. The IRS also offers a pathway to enhance the task offloading efficiency of the MEC network due to its considerable passive beamforming gain. By strategically situating IRSs near IoT devices, it is possible to effectively counteract significant signal attenuation due to distance or the non-line-of-sight (NLoS) condition, thereby appreciably extending the service coverage of MEC. This is integral to unlocking the full potential of MEC in offering high computational capabilities for future IoT networks [20].

In this paper, our focus is on the joint uplink (UL) resource allocation design. The key contributions are as follows:

- We study a novel joint radio resource allocation and edge offloading decision optimization in an IRS-assisted MEC network, where a multi-antenna AP receives information symbols from a set of MTC-enabled IoT users with finite block length transmissions. In particular, a resource allocation algorithm is designed to minimize the system's power consumption subject to peak transmit power feasibility and QoS constraints considering the interference.
- The formulated problem is non-convex and mixed integer non-linear programming (MINLP). To address this and find a suboptimal solution, we employ successive convex approximation (SCA) and a penalty-based approach to tackle the unit-modulus constraints because of passive reflecting elements at the IRS.
- The simulation results reveal that deploying IRS along with multi-antenna AP can realize low latency and high reliability in MEC-assisted systems with MTC IoT users.

This paper is organized as follows: Section II introduces the system and channel models. Section III formulates the proposed resource allocation problem. The resource allocation algorithm design policy is presented in Section IV. Section V evaluates the performance of the proposed schemes using computer simulations, and conclusions are drawn in Section VI.

II. System Model and Performance Metric

This section presents the system and channel models in the IRS-assisted frequency division multiple access (OFDMA) MEC system for IoT users with short packet lengths. In particular, we consider a single-cell multi-user UL communication that comprises an AP associated with a MEC server, equipped with N_{AP} antennas, as shown in Fig. 1. To provide edge computing services, the AP receives UL transmissions from K single antenna IoT users, indexed by k with the set of $\mathcal{K} = \{1, \ldots, K\}$, directly or via an IRS. The IRS is composed of M passive reflecting elements characterized by their phase shifts and amplitudes. Let's define $\mathbf{\Phi} = \text{diag}(\beta_1 e^{j\alpha_1}, \beta_2 e^{j\alpha_2}, ..., \beta_M e^{j\alpha_M})$ as the reflection-coefficients matrix at the IRS, where $\beta_m \in [0, 1]$ and $\alpha_m \in (0, 2\pi]$, $\forall m \in \{1, ..., M\}$ are the reflection amplitude[1] and phase shift

[1]For reflection efficiency maximization, the amplitudes of all passive elements are assumed to be one [21] i.e., $\beta_m = 1$, $\forall m$.

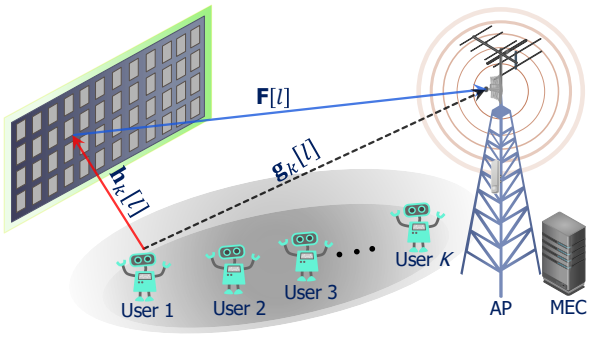

Fig. 1. Multi-user IRS-assisted MEC system with one AP, and K IoT users with finite block length transmissions. The single-antenna MTC-enabled users offload their tasks to a MEC server, either directly or via IRS, utilizing a multi-antenna AP.

of the m-th reflection coefficient at the IRS, respectively. The bandwidth is divided into N orthogonal sub-carriers indexed by $\mathcal{N} = \{1, \ldots, N\}$. The bandwidth of each sub-carrier is B_s, leading to a symbol duration of $T_s = \frac{1}{B_s}$. The UL frame is divided into L time slots indexed by $\mathcal{L} = \{1, 2, ..., L\}$. To obtain a performance upper bound, perfect channel state information (CSI) of the entire system is assumed to be available at the AP. We assume that the delay requirements of all users are known at the AP, and only users whose delay requirements can potentially be met in the current resource block—made up of L time slots and N sub-carriers—are admitted into the system. Each user has one computation task $(B_k,\ D_k)$ that needs to be processed, where B_k is the task (or input data or bitstream) size in bits and D_k is the delay deadline in time slots (also known as service delay).

A. Signal and Channel Models

Each IoT user transmits its own UL signal. The received signal at the AP in time slot l and subcarrier n is given by:

$$\mathbf{y}[l,n] = \sum_{k=1}^{K} \sqrt{p_k[l,n]}(\mathbf{F}[n]\mathbf{\Phi}\mathbf{h}_k[n] + \mathbf{g}_k[n])u_k[l,n] + \mathbf{z}[l,n], \tag{1}$$

where $\mathbf{h}_k[n] \in \mathbb{C}^{M\times 1}$ and $\mathbf{g}_k[n] \in \mathbb{C}^{N_{\text{AP}}\times 1}$ are the IRS-user and AP-user channel vectors of the k-th user. Also, $\mathbf{F}[n] \in \mathbb{C}^{N_{\text{AP}}\times M}$, $\mathbf{\Phi} \in \mathbb{C}^{M\times M}$, $u_k[l,n] \in \mathbb{C}$, and $p_k[l,n]$ are the AP-IRS channel matrix, phase shift matrix of the IRS, transmit symbol of user k on subcarrier n in time slot l, and the power of user k on subcarrier n and time slot l, respectively. Furthermore, $\mathbf{z}[l,n] \in \mathbb{C}^{N_{\text{AP}}\times 1}$ is the received noise vector at the AP with $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{N_{\text{AP}}})$, and we assume $\mathcal{E}\{|u_k[l,n]|^2\} = 1,\ \forall k, l, n$. The received signal vector on n-th subcarrier via adopting receive beamforming is given by:

$$\tilde{\mathbf{u}}[l,n] = \mathbb{\gamma}^H[l,n]\mathbf{y}[l,n],\ \forall l,\ \forall n, \tag{2}$$

where $\mathbb{\gamma}[l,n] \in \mathbb{C}^{N_{\text{AP}}\times K}$ is a matrix whose k-th columns are given by $[\mathbf{w}_k[n]] \in \mathbb{C}^{N_{\text{AP}}\times 1}, \forall k, l$. As a result, the signal-to-interference-plus-noise-ratio (SINR) of user k on subcarrier n in time slot l can be expressed as:

$$\gamma_k[l,n] = \frac{\|\mathbf{w}_k^H[n]\bar{\mathbf{h}}_k[n])\|^2 p_k[l,n]}{\sum_{j\neq k}^{K} \|\mathbf{w}_k^H[n]\bar{\mathbf{h}}_j[n])\|^2 p_j[l,n] + \tilde{\sigma}^2},\ \forall k, l, n, \tag{3}$$

where $\bar{\mathbf{h}}_k[n] = \mathbf{F}[n]\mathbf{\Phi}\mathbf{h}_k[n] + \mathbf{g}_k[n],\ \forall k, l, n$, and $\tilde{\sigma}^2 = \sigma^2 \|\mathbf{w}_k^H[l,n]\|^2,\ \forall k, l$.

B. Achievable Rate with Short Packet Transmission

By revisiting the SINR formula in (3), it is now possible to compute the achievable data rate for each IoT user. In MTC-enabled systems, low-latency wireless communication requires the use of finite and short blocklengths. The precise approximation for the achievable rate of each user can then be expressed as follows:

$$R_k(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) = F(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) - G(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k),\ \forall k, \quad (4)$$

where

$$F_k(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) = \sum_{l=1}^{L}\sum_{n=1}^{N} \log_2(1 + x_k[l,n]\gamma_k[l,n]),\ \forall k, \quad (5)$$

$$G_k(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) = Q^{-1}(\epsilon_k)\sqrt{\sum_{l=1}^{L}\sum_{n=1}^{N} x_k[l,n]V_k[l,n]},\ \forall k. \quad (6)$$

The subcarrier assignment indicators $x_k[l,n]$ in (5) and (6) determine whether subcarrier n in time slot l is assigned to user k. If assigned, $x_k[l,n] = 1$, otherwise $x_k[l,n] = 0$. The optimization variables $p_k[l,n], \forall l, n$ and $x_k[l,n], \forall l$, are collected in $\mathbf{p}_k$ and $\mathbf{x}_k$ respectively. Additionally, the decoding error is denoted by ϵ_k, the channel dispersion by $V_k[l,n]$, and $Q^{-1}(\cdot)$ represents the inverse of the Gaussian Q-function. The channel dispersion $V_k[l,n]$ is calculated as $V_k[l,n] = a^2\big(1 - (1 + \gamma_k[l,n])^{-2}\big)$, where $a = \log_2(e)$. To meet the user's delay requirements, all symbols of user k are assigned to the first $\mathfrak{d}_k$ time slots, i.e., $x_k[l,n] = 0, \forall l > \mathfrak{d}_k$.

C. Offloading Decision

Let's define s_k as the binary variable indicator for edge binary offloading decisions. In particular, if $s_k = 1$, user k offloads its processing task; otherwise, the task would be done locally. In order to guarantee the QoS of IoT user k, its service delay D_k should not exceed a given acceptable threshold, $T_{\max}$, in each time slot. Here, we assume that the time needed for data processing is short, and the response time delay is negligible. When IoT users choose to execute the data locally, the central processing unit (CPU) power consumption becomes the primary factor. This power consumption consists of dynamic power, short circuit power, and leakage power [22]. In [23], the authors demonstrated that at the optimal CPU frequency, the minimum power consumption of the CPU is directly proportional to $(\frac{B_k}{D_k})^{\mathfrak{c}}$, where $\mathfrak{c}$ denotes the power scaling factor. Therefore, we adopt the following model to estimate the power consumption of local execution:

$$\mathcal{E}_k^{\text{loc}} = (1 - s_k)\hat{\underline{\beta}}\frac{(B_k)^{\mathfrak{c}}}{(D_k)^{\mathfrak{c}}},\ \forall k, \quad (7)$$

where $\hat{\underline{\beta}}$ is a constant value that depends on the application parameter. On the other hand, IoT users have the option to offload their data to the edge server in the UL when necessary. Therefore, the IoT user's offloading transmission power for sending data to the edge server can be stated as follows:

$$\mathcal{E}_k^{\text{ofl}} = \sum_{l\in\mathcal{L}}\sum_{n\in\mathcal{N}} s_k x_k[l,n]p_k[l,n] + s_k p_{cir},\ \forall k, \quad (8)$$

where p_{cir} is the constant circuit power consumption during transmission. Consequently, the total power consumption of the system in the UL is represented by the sum of local and offloading power consumption, given by:

$$\mathcal{E}^{\text{total}} = \sum_{k\in\mathcal{K}} \left(\mathcal{E}_k^{\text{ofl}} + \mathcal{E}_k^{\text{loc}}\right). \quad (9)$$

III. Power Minimization Problem Formulation

In this section, we construct a joint problem of resource allocation and offloading decision with the goal of minimizing the total power consumption while satisfying the QoS requirements of MTC-enabled IoT users with short packet length transmission. In particular, we focus on optimizing the UL transmit power, the phase shift of the IRS, the subcarrier assignment, and offloading decision. To this end, the optimization problem is formulated as follows:

$$\mathcal{P}_1: \min_{\mathbf{p},\mathbf{\Phi},\mathbf{s},\mathbf{x}} \ \mathcal{E}^{\text{total}} \quad (10)$$

$$s.t.: \ D_k \le s_k T_{\max}, \qquad \forall k, \quad (10a)$$

$$\sum_{l=1}^{L}\sum_{n=1}^{N} x_k[l,n]p_k[l,n] \le s_k p_{k,\max}, \forall k, \quad (10b)$$

$$|\mathbf{\Phi}_{m,m}| = 1, \qquad \forall m, \quad (10c)$$

$$x_k[l,n] = 0, \qquad \forall l > \mathfrak{d}_k, \forall k, \forall n, \quad (10d)$$

$$\sum_{k=1}^{K} x_k[l,n] \le 1, \qquad \forall l, n, \quad (10e)$$

$$s_k \in \{0,1\}, \qquad \forall k, \quad (10f)$$

$$x_k[l,n] \in \{0,1\}, \qquad \forall k, l, n. \quad (10g)$$

In the problem formulation, the variables $\mathbf{p}$, $\mathbf{x}$, $\mathbf{s}$, and $\mathbf{\Phi}$ represent collections of optimization variables. These variables are utilized to optimize the system performance and make decisions related to power allocation, subcarrier assignment, offloading decisions, and phase shifts of the IRS elements, respectively. In $\mathcal{P}_1$, the constraint (10a) ensures that the delay in offloading tasks from each user to the edge computing does not exceed the given threshold $T_{\max}$. The constraint (10b) limits the transmit power of each user to be within the maximum power budget $p_{\max}$. The constraint (10c) imposes the unit modulus constraint on the IRS elements. The constraint (10d) is the user's delay requirements. The constraint (10e) indicates that each subcarrier can only be assigned to one user. Lastly, the constraints (10f) and (10g) represent the binary nature of the subcarrier assignment and offloading decision variables.

The problem $\mathcal{P}_1$ is a challenging non-convex mixed integer non-linear problem (MINLP) with interdependent optimization variables, non-convex phase shift constraints, and binary variables. Solving such non-convex optimization problems optimally is a complex task. However, we propose an efficient algorithm using the successive convex approximation (SCA) method. Moreover, we design a penalty-based approach to tackle the unit-modulus constraints because of passive reflecting elements at the IRS. Our solution is polynomial-time suboptimal yet computationally efficient.

IV. Joint Solution of the Optimization Problem

We begin our solution design framework by converting the delay threshold constraint (10a). This transformation is crucial as it allows for more tractable analysis. Since each user's data rate is determined by two key parameters, the size of its bitstream and the delay it experiences, the achievable data rate as per (4) corresponds to: $R_k(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) = \frac{B_k}{D_k}$. Accordingly, the equivalent form of $\mathcal{P}_1$ can be stated as:

$$\mathcal{P}_2: \min_{\mathbf{p},\mathbf{\Phi},\mathbf{s},\mathbf{x}} \ \mathcal{E}^{\text{total}} \tag{11}$$

$$s.t.: \ R_k(\mathbf{p}_k, \mathbf{\Phi}, \mathbf{x}_k) \geq s_k B_k, \quad \forall k, \tag{11a}$$

$$(10b) - (10g).$$

The transformed constraint, (11a), guarantees the offloading traffic for each user k is at least B_k bits. Secondly, to address the challenge posed by the multiplication of two binary variables in $\mathcal{P}_2$, we introduce s_k as a maximum value constraint on $x_k[l,n]$, that is, $x_k[l,n] \leq s_k$. This simplifies the term $s_k x_k[l,n]$ to $x_k[l,n]$. Hence, the optimization problem stated in equation $\mathcal{P}_2$ can be reformulated as:

$$\mathcal{P}_3: \min_{\mathbf{p},\mathbf{\Phi},\mathbf{s},\mathbf{x}} \ \bar{\mathcal{E}}^{\text{total}} = \sum_{k\in\mathcal{K}} \left(\bar{\mathcal{E}}_k^{\text{ofl}} + \mathcal{E}_k^{\text{loc}}\right) \tag{12}$$

$$s.t.: \ \sum_{l=1}^{L}\sum_{n=1}^{N} p_k[l,n] \leq p_{k,\max}, \forall k, \tag{12a}$$

$$\sum_{l=1}^{L} p_k[l,n] \leq s_k p_{k,\max}, \forall k, n, \tag{12b}$$

$$x_k[l,n] \leq s_k, \forall k, l, n, \tag{12c}$$

$$(10b) - (10g), (11a),$$

where $\bar{\mathcal{E}}_k^{\text{ofl}} = \sum_{l\in\mathcal{L}}\sum_{n\in\mathcal{N}} x_k[l,n]p_k[l,n] + s_k p_{cir}, \ \forall k$. To overcome the non-convexity of multiplying $x_k[l,n]$ and $p_k[l,n]$, we introduce a new variable $\tilde{p}_k[l,n] = x_k[l,n]p_k[l,n]$. Using the big-M formulation [24], we add new constraints into $\mathcal{P}_3$ to manage this non-convex term as follows:

$$\mathcal{P}_4: \min_{\mathbf{p},\tilde{\mathbf{p}},\mathbf{\Phi},\mathbf{s},\mathbf{x}} \ \tilde{\mathcal{E}}^{\text{total}} = \sum_{k\in\mathcal{K}} \left(\tilde{\mathcal{E}}_k^{\text{ofl}} + \mathcal{E}_k^{\text{loc}}\right) \tag{13}$$

$$s.t.: \ \tilde{R}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k) \geq s_k B_k, \quad \forall k, \tag{13a}$$

$$\sum_{l=1}^{L}\sum_{n=1}^{N} \tilde{p}_k[l,n] \leq s_k p_{k,\max}, \quad \forall k, \tag{13b}$$

$$\tilde{p}_k[l,n] \leq x_k[l,n] p_{k,\max}, \quad \forall k, n, l, \tag{13c}$$

$$\tilde{p}_k[l,n] \leq p_k[l,n], \quad \forall k, n, l, \tag{13d}$$

$$\tilde{p}_k[l,n] \geq p_k[l,n] - (1 - x_k[l,n])p_{k,\max}, \forall k, n, l, \tag{13e}$$

$$\tilde{p}_k[l,n] \geq 0, \quad \forall k, n, l, \tag{13f}$$

$$(10c) - (10g), (12a) - (12c),$$

where $\tilde{\mathcal{E}}_k^{\text{ofl}} = \sum_{l\in\mathcal{L}}\sum_{n\in\mathcal{N}} \tilde{p}_k[l,n] + s_k p_{cir}, \ \forall k$. Moreover, $\tilde{R}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k) = \tilde{F}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k) - \tilde{G}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k)$, where

$$\tilde{\gamma}_k[l,n] = \frac{\|\mathbf{w}_k^H[n]\bar{\mathbf{h}}_k[n])\|^2 \tilde{p}_k[l,n]}{\sum_{j\neq k}^{K} \|\mathbf{w}_k^H[n]\bar{\mathbf{h}}_j[n])\|^2 \tilde{p}_j[l,n] + \tilde{\sigma}^2}, \forall k, l, n, \tag{14}$$

$$\tilde{F}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k) = \sum_{l=1}^{L}\sum_{n=1}^{N} \log_2(1 + \tilde{\gamma}_k[l,n]), \ \forall k, \tag{15}$$

$$\tilde{G}_k(\mathbf{p}_k, \tilde{\mathbf{p}}_k, \mathbf{\Phi}, \mathbf{x}_k) = aQ^{-1}(\epsilon_k)\sqrt{\sum_{l=1}^{L}\sum_{n=1}^{N}(1 - (1+\tilde{\gamma}_k[l,n])^{-2})}, \forall k. \tag{16}$$

Furthermore, $\tilde{\mathbf{p}}_k$ represents the collection of optimization variables $\tilde{p}_k[l,n], \forall l, n$. Next, we relax the integer variables by converting them into continuous variables within the range of zero to one. We then impose the following constraints on the optimization problem to define the feasible regions:

$$\mathcal{P}_5: \min_{\mathbf{p},\tilde{\mathbf{p}},\mathbf{\Phi},\mathbf{s},\mathbf{x}} \ \tilde{\mathcal{E}}^{\text{total}} + \lambda_1\left(\sum_{k=1}^{K} s_k - s_k^2\right) + \lambda_2\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{n=1}^{N} x_k[l,n] - x_k[l,n]^2\right) \tag{17}$$

$$s.t.: \ 0 \leqslant s_k \leqslant 1, \quad \forall k, l, \tag{17a}$$

$$0 \leqslant x_k[l,n] \leqslant 1, \quad \forall k, l, n, \tag{17b}$$

$$(10c) - (10e), (12a) - (12c), (13a) - (13f).$$

where λ_1 and λ_2 are penalty factors that need to be greater than one. We now transform the SINR function in (14) into a mathematically tractable form to optimize the phase shifts. By adopting semidefinite programming (SDP), we have:

$$\|\mathbf{w}_k^H[n]\bar{\mathbf{h}}_k[n]\|^2 = \text{Tr}\big(\mathbf{U}_k[n]\mathbf{\Upsilon}\mathbf{U}_k^H[n]\mathbf{W}_k[n]\big), \tag{18}$$

where $\mathbf{U}_k[n] = \left[\left(\mathbf{F}^H[n]\text{diag}(\mathbf{h}_k^{\text{H}}[n])\right)^{\text{T}} \ \mathbf{g}_k^*[n]\right]^{\text{T}}$, $\mathbf{W}_k[n] = \mathbf{w}_k[l,n]\mathbf{w}_k^H[l,n]$, $\mathbf{\Upsilon} = \boldsymbol{\varpi}\boldsymbol{\varpi}^H \in \mathbb{C}^{(M+1)\times(M+1)}$, $\boldsymbol{\varpi} = [\boldsymbol{\varrho}^T \ \varkappa]^T \in \mathbb{C}^{(M+1)\times 1}$, where $\varkappa \in \mathbb{C}$ is a dummy variable with $|\varkappa|^2 = 1$, and $\boldsymbol{\varrho} = [e^{j\alpha_1}, e^{j\alpha_2}, \ldots, e^{j\alpha_M}]^H \in \mathbb{C}^{M\times 1}$. To disentangle the complexity and facilitate the solution of $\mathcal{P}_5$, we introduce a set of auxiliary variables $\chi_k[l,n], \forall k, l, n$, to establish a lower bound on the SINR as defined in equation (14), that is:

$$0 \leq \chi_k[l,n] \leq \tilde{\gamma}_k[l,n] \triangleq \frac{\mathcal{C}_k[l,n]}{\mathcal{D}_k[l,n]}, \forall k, l, n, \tag{19}$$

where

$$\mathcal{C}_k[l,n] = \text{Tr}\big(\mathbf{Z}_k[n]\mathbf{W}_k[n]\big)\tilde{p}_k[l,n], \tag{20}$$

$$\mathcal{D}_k[l,n] = \sum_{j\neq k}^{K} \text{Tr}\big(\mathbf{Z}_j[n]\mathbf{W}_k[n]\big)\tilde{p}_j[l,n] + \tilde{\sigma}^2, \tag{21}$$

with $\mathbf{Z}_q[n] = \mathbf{U}_q[n]\mathbf{\Upsilon}\mathbf{U}_q^H[n], \forall q = \{k, j\}$. Consequently, the achievable rate in (13a) can be restated as $\tilde{R}_k(\boldsymbol{\chi}_k) = \tilde{F}_k(\boldsymbol{\chi}_k) - \tilde{G}_k(\boldsymbol{\chi}_k)$, where $\boldsymbol{\chi}_k$ is the collection of optimization variables $\chi_k[l,n], \forall l, n$. We now define slack optimization variables $\mathcal{I}_k[l,n], \forall k, l, n$ to set an upper bound on the denominator of (19). By incorporating these slack optimization variables, we can regulate and control the SINR, leading to the following reformulation:

$$\chi_k[l,n]\mathcal{I}_k[l,n] \leq \mathcal{C}_k[l,n], \ \forall k, l, n, \tag{22}$$

$$\mathcal{I}_k[l,n] \geq \mathcal{D}_k[l,n], \quad \forall k, l, n, \tag{23}$$

where $\mathcal{I}_k[l,n]$ denotes the k-th user interference on time slot l and subcarrier n. This method of constraint manipulation is a common approach used in optimization theory and is vital in our case as it simplifies the complex function, transforming it into a more manageable form. By referring to the objective function of $\mathcal{P}_5$ as $\ddot{\mathcal{E}}^{\text{total}}$, the revised optimization problem is

formulated as follows:

$$\mathcal{P}_6: \min_{\mathbf{p},\tilde{\mathbf{p}},\mathbf{\Phi},\mathbf{s},\mathbf{x},\boldsymbol{\chi},\boldsymbol{\Upsilon}} \ \ddot{\mathcal{E}}^{\text{total}} \tag{24}$$

$$s.t.: \tilde{F}_k(\boldsymbol{\chi}_k) - \tilde{G}_k(\boldsymbol{\chi}_k) \geq s_k B_k, \quad \forall k, \tag{24a}$$

$$\text{diag}(\boldsymbol{\Upsilon}) = \mathbf{1}_{\text{M}+1}, \tag{24b}$$

$$\boldsymbol{\Upsilon} \succeq 0, \tag{24c}$$

$$\text{rank}(\boldsymbol{\Upsilon}) \leq 1, \tag{24d}$$

$$(10\text{c}) - (10\text{e}), (12\text{a}) - (12\text{c}), (13\text{a}) - (13\text{f}),$$

$$(17\text{a}), (17\text{b}), (22), (23).$$

The optimization problem $\mathcal{P}_6$ remains non-convex due to the non-convex nature of (22), (23), and the rank constraint (24d). We also observe that constraints (22) and (23) fall into the category of difference of convex (DC) problems [7], [18], [24], [25]. In addition, the bilinear term $\chi_k[l,n]\mathcal{I}_k[l,n]$ on the left-hand side of (22) is non-convex, which poses an extra challenge in designing an efficient resource allocation algorithm. However, this product can also be expressed as the difference of two convex functions, leading to a DC representation of (22) and (23) as follows:

$$\varsigma_1(\chi_k[l,n], \mathcal{I}_k[l,n]) - \varsigma_2(\chi_k[l,n], \mathcal{I}_k[l,n]) \leq \big(\varsigma_3(\boldsymbol{\Upsilon}, \tilde{p}_k[l,n]) - \varsigma_4(\boldsymbol{\Upsilon}, \tilde{p}_k[l,n])\big), \tag{25}$$

$$\sum_{j\neq k}^{K} \big(\varsigma_5(\boldsymbol{\Upsilon}, \tilde{p}_j[l,n]) - \varsigma_6(\boldsymbol{\Upsilon}, \tilde{p}_j[l,n])\big) + \tilde{\sigma}^2 \leq I_k[l,n], \tag{26}$$

where

$$\varsigma_1(\chi_k[l,n], \mathcal{I}_k[l,n]) = 0.5(\chi_k[l,n] + \mathcal{I}_k[l,n])^2, \tag{27}$$

$$\varsigma_2(\chi_k[l,n], \mathcal{I}_k[l,n]) = 0.5(\chi_k[l,n])^2 + 0.5(\mathcal{I}_k[l,n])^2, \tag{28}$$

$$\varsigma_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n]) = 0.5\big(\tilde{p}_\vartheta[l,n] + \text{Tr}(\mathbf{Z}_j[n]\mathbf{W}_k[n])\big)^2, \quad \forall\{\zeta,\vartheta\} = \{\{3,k\},\{5,j\}\}, \tag{29}$$

$$\varsigma_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n]) = 0.5(\tilde{p}_\vartheta[l,n])^2 + 0.5\big(\text{Tr}(\mathbf{Z}_j[n]\mathbf{W}_k[n])\big)^2, \quad \forall\{\zeta,\vartheta\} = \{\{4,k\},\{6,j\}\}. \tag{30}$$

Both sides of (25) and the left-hand side of (26) are not convex. To address the non-convexity of the left-hand side of (25), we apply the SCA technique, which involves using a first-order Taylor expansion to obtain a convex approximation of the non-convex terms as follows:

$$\bar{\varsigma}_2(\chi_k[l,n], \chi_k^{(i)}[l,n], \mathcal{I}_k[l,n], \mathcal{I}_k^{(i)}[l,n]) = 0.5(\chi_k^{(i)}[l,n])^2 + \chi_k^{(i)}[l,n](\chi_k[l,n] - \chi_k^{(i)}[l,n]) + 0.5(\mathcal{I}_k^{(i)}[l,n])^2 + \mathcal{I}_k^{(i)}[l,n](\mathcal{I}_k[l,n] - \mathcal{I}_k^{(i)}[l,n]), \forall k,l,n. \tag{31}$$

Similarly, to address the non-convexity of the right-hand side of (25) and the left-hand side of (26), we also employ the SCA technique. Thus, we can approximate these non-convex terms as follows:

$$\tilde{\varsigma}_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n], \boldsymbol{\Upsilon}^{(i)}, \tilde{p}_\vartheta^{(i)}[l,n]) = \varsigma_\zeta(\boldsymbol{\Upsilon}^{(i)}, \tilde{p}_\vartheta^{(i)}[l,n]) + \text{Tr}(\nabla_{\boldsymbol{\Upsilon}}(\varsigma_\zeta(\boldsymbol{\Upsilon}^{(i)}, \tilde{p}_\vartheta^{(i)}[l,n])^H(\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^{(i)})) + \text{Tr}(\nabla_{\tilde{p}_\vartheta}(\varsigma_\zeta(\boldsymbol{\Upsilon}^{(i)}, \tilde{p}_\vartheta^{(i)}[l,n])^H(\tilde{p}_\vartheta[l,n] - \tilde{p}_\vartheta^{(i)}[l,n])), \forall\{\zeta,\vartheta\} = \{\{4,k\},\{6,j\}\}, \tag{32}$$

where $\nabla_{\boldsymbol{\Upsilon}}(\varsigma_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n])$ and $\nabla_{\tilde{p}_\vartheta}(\varsigma_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n])$ are the gradients of $\varsigma_\zeta(\boldsymbol{\Upsilon}, \tilde{p}_\vartheta[l,n])$, (30), with respect to $\boldsymbol{\Upsilon}$ and $\tilde{p}_\vartheta$, respectively. Therefore, (25) and (26) can be approximated as follows:

$$\varsigma_1(\chi_k[l,n], \mathcal{I}_k[l,n]) - \bar{\varsigma}_2(\chi_k[l,n], \chi_k^{(i)}[l,n], \mathcal{I}_k[l,n], \mathcal{I}_k^{(i)}[l,n]) \leq \big(\varsigma_3(\boldsymbol{\Upsilon}, \tilde{p}_k[l,n]) - \tilde{\varsigma}_4(\boldsymbol{\Upsilon}, \tilde{p}_k[l,n], \boldsymbol{\Upsilon}^{(i)}, \tilde{p}_k^{(i)}[l,n])\big), \tag{33}$$

$$\sum_{j\neq k}^{K} \big(\varsigma_5(\boldsymbol{\Upsilon}, \tilde{p}_j[l,n]) - \tilde{\varsigma}_6(\boldsymbol{\Upsilon}, \tilde{p}_j[l,n], \boldsymbol{\Upsilon}^{(i)}, \tilde{p}_j^{(i)}[l,n])\big) + \tilde{\sigma}^2 \leq I_k[l,n], \forall k,l,n. \tag{34}$$

$\mathcal{P}_6$ is still not convex. Finally, the convexity of $\mathcal{P}_6$ hinges on the rank of $\boldsymbol{\Upsilon}$. Typically, $\mathcal{P}_6$ yields solutions with a rank higher than one. To prevail over this last challenge, we reformulate constraint (24d) utilizing the DC method, resulting in the following expression:

$$||\boldsymbol{\Upsilon}||^* - ||\boldsymbol{\Upsilon}||^2 \leq 0. \tag{35}$$

Note that $||\boldsymbol{\Upsilon}||^* = \sum_i \tau_i \geq ||\boldsymbol{\Upsilon}||^2 = \max_i\{\tau_i\}$ holds for any given $\boldsymbol{\Upsilon}$, where τ_i is the i-th singular value of $\boldsymbol{\Upsilon}$. The equality holds if and only if $\boldsymbol{\Upsilon}$ achieves rank one i.e., $\text{rank}(\boldsymbol{\Upsilon}) = 1$ [7]. Now, we take the first-order Taylor approximation of $||\boldsymbol{\Upsilon}||^2$ as:

$$||\boldsymbol{\Upsilon}||^2 \geq \overbrace{||\boldsymbol{\Upsilon}^{(t)}||^2 + \text{Tr}\big(\lambda_{\max}(\boldsymbol{\Upsilon}^{(t)})\lambda_{\max}^H(\boldsymbol{\Upsilon}^{(t)})(\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^t)\big)}^{=\kappa(\boldsymbol{\Upsilon})}, \tag{36}$$

where $\lambda_{\max}(\cdot)$ returns the largest eigenvalue of $\boldsymbol{\Upsilon}^{(t)}$. By utilizing (36), we can obtain a convex approximation for (35), expressed as $\tilde{\kappa}^t(\boldsymbol{\Upsilon}) \triangleq ||\boldsymbol{\Upsilon}||_* - \kappa(\boldsymbol{\Upsilon}) \leq 0$. Finally, the optimization problem is formulated by adding $\tilde{\kappa}^t(\boldsymbol{\Upsilon})$ to the objective function of $\mathcal{P}_6$ with a penalty factor $\delta \gg 1$ to penalize non-rank-one matrices, as follows:

$$\mathcal{P}_7: \min_{\mathbf{p},\tilde{\mathbf{p}},\mathbf{\Phi},\mathbf{s},\mathbf{x},\boldsymbol{\chi},\boldsymbol{\Upsilon}} \ \ddot{\mathcal{E}}^{\text{total}} + \delta(\tilde{\kappa}^t(\boldsymbol{\Upsilon})) \tag{37}$$

$$(10\text{c}) - (10\text{e}), (12\text{a}) - (12\text{c}), (13\text{a}) - (13\text{f}),$$

$$(17\text{a}), (17\text{b}), (24\text{a}) - (24\text{c}), (33), (34).$$

The optimization problem $\mathcal{P}_7$ can be effectively solved by utilizing well-established convex optimization packages like CVX [7], [18], [24], [25]. We outline our proposed joint resource allocation algorithm in **Algorithm 1**.

Algorithm 1 Proposed Joint Iterative SCA Algorithm

Input: Set iteration index $i = 1$, and maximum number of iteration $\mathcal{T}_{\max}$, randomly initialize $\mathbf{p}^0$, $\tilde{\mathbf{p}}^0$, $\mathbf{\Phi}^0$, $\mathbf{s}^0$, $\mathbf{x}^0$, $\boldsymbol{\chi}^0$, $\boldsymbol{\Upsilon}^0$, and penalty factors $[\lambda_1, \lambda_2, \delta]^T \succ \mathbf{1}_3$

1: **repeat**
2: Calculate (31) and (32)
3: Solve $\mathcal{P}_7$ for given $\mathbf{p}^{(i)}$, $\tilde{\mathbf{p}}^{(i)}$, $\mathbf{\Phi}^{(i)}$, $\mathbf{s}^{(i)}$, $\mathbf{x}^{(i)}$, $\boldsymbol{\chi}^{(i)}$, and $\boldsymbol{\Upsilon}^{(i)}$, and retain the intermediate solution
4: Set $i = i + 1$ and $\mathbf{p}^{(i)} = \mathbf{p}^*$, $\tilde{\mathbf{p}}^{(i)} = \tilde{\mathbf{p}}^*$, $\mathbf{\Phi}^{(i)} = \mathbf{\Phi}^*$, $\mathbf{s}^{(i)} = \mathbf{s}^*$, $\mathbf{x}^{(i)} = \mathbf{x}^*$, $\boldsymbol{\chi}^{(i)} = \boldsymbol{\chi}^*$, and $\boldsymbol{\Upsilon}^{(i)} = \boldsymbol{\Upsilon}^*$
5: **until** $i = \mathcal{T}_{\max}$
6: **return** $\mathbf{p}^*$, $\tilde{\mathbf{p}}^*$, $\mathbf{\Phi}^*$, $\mathbf{s}^*$, $\mathbf{x}^*$, $\boldsymbol{\chi}^*$, $\boldsymbol{\Upsilon}^*$

V. Performance Evaluation

In this section, we present simulation results to verify the performance of the proposed IRS-assisted MEC system using the simulation parameters provided in **Table I**, unless otherwise stated. We assume the center of the network is the point $(0,0)$ m and the AP is located at $(0,-100)$ m while the IRS is located at $(50,0)$. Moreover, five MTC-enabled

Table I. Simulation Parameters

Parameter	Value
Total number of reflecting elements M	50
Total number of UL time slots L	4
Total number of subcarriers N	32
Bandwidth of each sub-carrier	30 kHz
Noise power density	-174 dBm/Hz
Maximum transmit power of each user, $p_{k,\max}$	23 dBm
Circuit power consumption of user k, p_{cir}	50 mW
Packet decoding error probability, ϵ_k	10^{-6}
Number of bits per packet (bitstream size), B_k	160 bits

IoT users ($K = 5$) are randomly distributed inside a circle with a radius of 4 m, and the center of the circle is the point $(25, 0)$ m. The path-loss model, based on distance, is defined as $L(d) = a_0(d/d_0)^{-\xi}$, where $a_0 = -30$ dB signifies the signal attenuation at a reference distance of $d_0 = 1$ m, d represents the link distance, and ξ is the path-loss exponent. Path-loss exponents for the AP-IRS, IRS-user, and AP-user links are respectively set at 2.2, 2.2, and 3.4 [7]. We presume that all links exhibit Rician fading, characterized by a Rician factor of 3 dB [7], [19]. For the small-scale fading, we assume Rayleigh fading for the AP-user channel and Rician fading with a Rician factor of 10 for the AP-IRS and IRS-user channels [7], [19].

A. Performance Bound and Benchmark Schemes

We assess the proposed resource allocation algorithm's efficiency by comparing it to the following baseline schemes:

- **Lower bound**: To obtain a lower bound on the system performance, Shannon's capacity formula is adopted in problem $\mathcal{P}_1$, i.e., $V_k[l, n], \forall k$, (6), is set to zero. The resulting optimization problem is solved using a modified version of the proposed algorithm.
- **Method A**: This is the proposed **Algorithm** 1.
- **Method B**: In this scheme, we adopt random phase shifts for the IRS elements and optimize the users' power allocation and offloading decisions.
- **Method C**: In this approach, we maintain a fixed sub-carrier allocation, based on [7] and [24], for offloading while optimizing other variables using SCA.
- **Method D**: We remove the IRS from the system in this scheme. We consider the UL power allocation and passive beamforming based solely on the direct link between the AP and the user.

B. Simulation Results

Fig. 2 illustrates the relationship between system power consumption and packet error probability, highlighting the impact of acceptable error rate and joint resource optimization on power consumption. The power consumption is observed to be a monotonically decreasing function of the packet error probability. This behavior is attributed to the fact that the complementary error function, used in the normal approximation for the data rate function, i.e., (4), exhibits a monotonically decreasing trend with respect to the packet error probability. Consequently, as the packet error probability increases, the influence of the dissipation part in the normal approximation

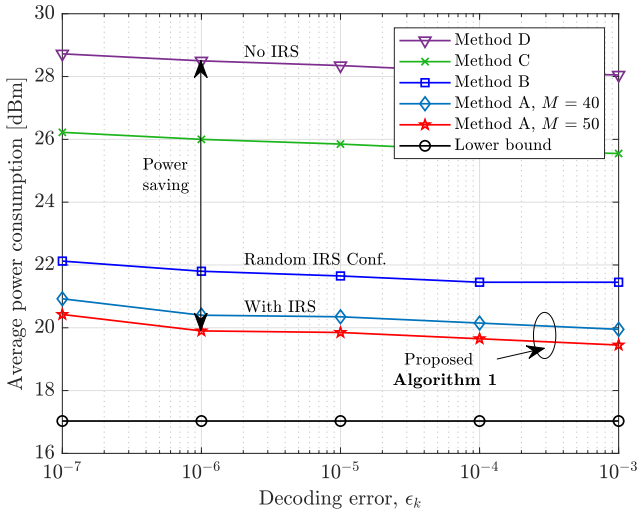

Fig. 2. Average power consumption [dBm] vs. decoding error probability.

diminishes. It is evident that a higher acceptable packet error probability leads to a lower transmit power required to meet the latency constraints of MTC-enabled IoT users. Moreover, this figure also shows that for the lower bound on performance, the power consumption is independent of the packet error probability (lower bound). This is because of the zero packet error probability assumption on Shannon's capacity formula (that is, $V_k[l, n], \forall k$, (6)). The gap between the lower bound and our proposed algorithm (Method A) represents the trade-off required to meet the stringent delay and reliability requirements of ultra-reliable low-latency communication with short packet transmission. The presence of IRSs in the system leads to substantial power savings, resulting in improved energy efficiency. However, it is important to consider the computational costs and investment required for deploying IRSs. Increasing the number of reflecting elements (M) enhances the passive beamforming gain, thereby reducing the transmit power of IoT devices and facilitating efficient offloading. The performance of the scheme utilizing random IRS beamforming (Method B) is inferior to that of a system utilizing an optimal beamforming vector (Method A), but it outperforms a system with fixed sub-carrier allocation (Method C). It is important to note that deploying IRSs plays a crucial role in maximizing the capabilities of MEC servers. By dynamically modifying the wireless propagation environment in real time, IRSs help ensure that users are not forced to allocate more power due to poor channel conditions. This allows for efficient offloading of user tasks to edge servers rather than having to compute the tasks locally. Furthermore, deploying IRSs helps guarantee that transmissions are completed within the desired delay, meeting the latency requirements of the system. This highlights the significance of IRSs in optimizing system performance and enabling effective task offloading in MEC environments.

In Fig. 3, the power consumption of the system is shown for different task sizes (a.k.a. bitstream sizes). As task sizes increase, power consumption increases in all schemes due to the need for higher SINRs and increased transmit power. The presence of IRSs improves SINRs by providing additional

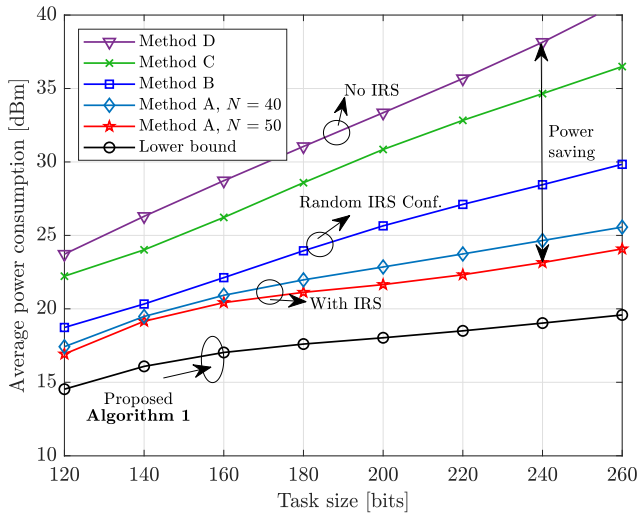

Fig. 3. Average power consumption [dBm] vs. the task size [bits].

LoS links, enabling the system to handle larger task sizes (method A) compared to systems without IRSs (method D). The proposed algorithm outperforms non-optimal subcarrier allocation and random IRS configuration by leveraging the improved SINRs (methods B and C). Moreover, the proposed method effectively reduces power consumption by optimizing offloading, subchannel allocation, and transmission power. The results emphasize the importance of choosing between offloading and local computing, especially when processing larger data sets.

VI. Conclusions

This paper studied the resource allocation algorithm design for a UL multi-user IRS-aided MEC system. In order to meet the strict requirements for end-to-end transmission delay and reliability inherent to MTC-enabled IoT users, we investigated a joint resource allocation and offloading decision scheme that considers short packet transmission. The IRS was deployed to enhance the communication channel and to increase reliability by providing virtual LoS links. We formulated an optimization problem for the minimization of average system power consumption subject to QoS constraints. The resulting problem was a non-convex MINLP, which posed significant challenges to finding a solution. An efficient, low-complexity algorithm was designed, utilizing SCA and an iterative rank minimization method, allowing it to find a local optimum. The simulation results highlighted the efficacy of our proposed algorithm and underscored the practical value of IRS in MEC systems for expanding coverage and assisting multiple energy-constrained devices with binary task offloading.

References

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, pp. 22–32, Feb. 2014.

[2] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, pp. 36–42, May 2016.

[3] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges," *IEEE Netw.*, vol. 35, pp. 158–167, Sep./Oct. 2021.

[4] L. Shen, N. Wang, Z. Zhu, W. Xu, Y. Li, X. Mu, and L. Cai, "UAV-Enabled data collection over clustered machine-type communication networks: AEM modeling and trajectory planning," *IEEE Trans. Veh. Technol.*, vol. 71, pp. 10016–10032, Sep. 2022.

[5] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, pp. 45–55, Nov. 2014.

[6] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, pp. 119–125, Dec. 2018.

[7] J. Jalali, A. Khalili, A. Rezaei, R. Berkvens, M. Weyn, and J. Famaey, "IRS-based energy efficiency and admission control maximization for IoT users with short packet lengths," *IEEE Trans. Veh. Technol.*, pp. 1–6, Apr. 2023.

[8] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, pp. 1628–1656, 3rd Quart. 2017.

[9] Z. A. El Houda, B. Brik, A. Ksentini, and L. Khoukhi, "A MEC-Based architecture to secure IoT applications using federated deep learning," *IEEE Internet Things J. Mag.*, vol. 6, pp. 60–63, Mar. 2023.

[10] J. Liang, H. Xing, F. Wang, and V. K. N. Lau, "Joint task offloading and cache placement for energy-efficient mobile edge computing systems," *IEEE Wirel. Commun. Lett.*, vol. 12, pp. 694–698, Apr. 2023.

[11] Q. Gan, G. Li, W. He, Y. Zhao, Y. Song, and C. Xu, "Delay-minimization offloading scheme in multi-server MEC networks," *IEEE Wirel. Commun.Lett.*, vol. 12, pp. 1071–1075, Jun. 2023.

[12] L. Li and P. Fan, "Latency and task loss probability for NOMA assisted MEC in mobility-aware vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 72, pp. 6891–6895, May. 2023.

[13] H. Yu and K.-W. Chin, "Maximizing sensing and computation rate in Ad Hoc energy harvesting IoT networks," *IEEE Internet Things J.*, vol. 10, pp. 5434–5446, Mar. 2023.

[14] L. Shi, X. Chu, H. Sun, and G. Lu, "Wireless-powered OFDMA-MEC networks with hybrid active–passive communications," *IEEE Internet Things J.*, vol. 10, pp. 10484–10496, Jun. 2023.

[15] M. Wu, W. Qi, J. Park, P. Lin, L. Guo, and I. Lee, "Residual energy maximization for wireless powered mobile edge computing systems with mixed-offloading," *IEEE Trans. Veh. Tech*, vol. 71, pp. 4523–4528, Apr. 2022.

[16] X. Qi, M. Peng, and H. Zhang, "Joint mmWave beamforming and resource allocation in NOMA-MEC network for Internet of Things," *IEEE Trans. Veh. Technol.*, vol. 72, pp. 4969–4980, Apr. 2023.

[17] Q. Zhang, Y. Wang, H. Li, S. Hou, and Z. Song, "Resource allocation for energy efficient STAR-RIS aided MEC systems," *IEEE Wirel. Commun. Lett.*, vol. 12, pp. 610–614, Apr. 2023.

[18] J. Jalali, A. Rezaei, A. Khalili, and J. Famaey, "Power-efficient joint resource allocation and decoding error probability for multiuser downlink MISO with finite block length codes," in *2022 25th Int. Symp. Wirel. Personal Multimed. Commun. (WPMC)*, pp. 232–237, Oct. 2022.

[19] J. Jalali, A. Khalili, A. Rezaei, and J. Famaey, "Is active IRS useful for mmwave wireless networks or not?," in *2023 Int. Conf. Computing, Netw. Commun. (ICNC)*, pp. 377–382, Feb. 2023.

[20] G. Chen, Q. Wu, R. Liu, J. Wu, and C. Fang, "IRS aided MEC systems with binary offloading: A unified framework for dynamic IRS beamforming," *IEEE J. Sel. Areas Commun.*, vol. 41, pp. 349–365, Feb. 2023.

[21] X. Yu, D. Xu, and R. Schober, "Optimal beamforming for MISO communications via intelligent reflecting surfaces," in *Proc. IEEE SPAWC*, pp. 1–5, 2020.

[22] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, pp. 4268–4282, Oct. 2016.

[23] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wirel. Commun.*, vol. 12, pp. 4569–4581, Sep. 2013.

[24] J. Jalali and A. Khalili, "Optimal resource allocation for MC-NOMA in SWIPT-Enabled networks," *IEEE Commun. Lett.*, vol. 24, pp. 2250–2254, Oct. 2020.

[25] J. Jalali, A. Khalili, A. Rezaei, J. Famaey, and W. Saad, "Power-efficient antenna switching and beamforming design for multi-user swipt with non-linear energy harvesting," in *2023 IEEE 20th Consumer Commun. Netw. Conf. (CCNC)*, pp. 746–751, Jan. 2023.