Joint Sensing, Communications, and Computing Design for 6G URLLC Service-Oriented MEC Networks

Dang Van Huynh, *Member, IEEE,* Saeed R. Khosravirad, *Senior Member, IEEE,*
Simon Cotton, *Senior Member, IEEE,* Thang X. Vu, *Senior Member, IEEE,*
Octavia A. Dobre, *Fellow, IEEE,* Hyundong Shin, *Fellow, IEEE,* and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—The convergence of advanced communication technologies and powerful computing architecture has unlocked a plethora of opportunities for Internet-of-Things applications. To fully realise this potential, a synergistic design encompassing sensing, computing, and communication is crucial. This paper investigates these critical technologies to facilitate service-oriented systems by minimising end-to-end latency and the number of deployed services at edge servers in mobile edge computing, all within the confines of stringent ultra-reliable and low-latency communication requirements and system budget constraints. The addressed optimisation problem takes into account various variables such as service placement strategies, task offloading portions, and bandwidth allocation. Simulation results validate the effectiveness of our solution and highlight the impact of key parameters on system performance.**

***Index Terms*—Integrated sensing and communications, mobile edge computing, service-oriented networks, task offloading, ultra-reliable and low latency communications.**

I. Introduction

Mobile edge computing (MEC) has emerged as a promising technology poised to unlock the full potential of future computing systems. With its distinctive computing architecture, MEC offers robust support for a multitude of delay-sensitive applications, ranging from industrial control and augmented reality/virtual reality (AR/VR) to intelligent transport systems [2]. Leveraging the MEC paradigm, constrained devices such as actuators and sensors can offload computational tasks to edge servers (ESs) via task offloading mechanisms, thereby reducing processing time. Moreover, ESs can serve as efficient data caches, diminishing the volume of information that needs to be transmitted [3]. However, despite these numerous benefits, MEC presents various challenges which require resolution, including offloading decisions, task caching, and the joint allocation of computation and communication resources [4], [5]. Consequently, MEC has garnered increasing attention from researchers in both academia and industry [4], [6].

Ultra-reliable and low-latency communications (URLLC) technology stands out among advanced communication techniques as a pivotal enabler for 5G and beyond mission-critical applications [7]. URLLC-based transmissions play a crucial role in various emerging services with stringent reliability and delay requirements, including motion control in industry, telesurgery in healthcare, and the tactile Internet [8]–[10]. Existing studies on URLLC have primarily centred on resource allocation, beamforming design, and minimising decoding error probability based on short-packet transmissions [11], [12]. These studies have delved into the intricate trade-off between transmission rate and reliability in URLLC-based mission-critical applications. In recent years, expanded research on URLLC in conjunction with other emerging technologies such as multi-tier computing [13]–[15] has opened up new implementation opportunities in real-world deployments [16]. This research direction holds tremendous potential for realising the vision of URLLC-aided MEC in future wireless networks [17].

Task offloading and service placement are two fundamental pillars of MEC. Service placement entails configuring the service platform, encompassing services, software, libraries, storage, and databases at Edge Servers (ESs) to cater to specific tasks [18]. This involves strategically positioning these components to ensure they are available where they are most needed. In practice, ESs can only accommodate a limited number of services due to resource constraints, necessitating careful management of computational and storage resources. Optimal service placement strategies not only enhance system performance and reduce deployment costs but also improve scalability, ensure high availability, and mitigate security risks by minimizing potential attack surfaces and ensuring critical services are properly isolated [19]. Considering service placement in the task offloading problem effectively bridges the

D. V. Huynh, O. A. Dobre are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X5, Canada (e-mail: {vdhuynh, odobre}@mun.ca).

S. R. Khosravirad is with Nokia Bell Labs, Murray Hill, NJ 07964 USA (e-mail: saeed.khosravirad@nokia-bell-labs.com).

S. Cotton is with the Centre for Wireless Innovation (CWI), School of Electronics, Electrical Engineering and Computer Science, ECIT Institute, Queen's University Belfast, BT3 9DT, Belfast, U.K (email: simon.cotton@qub.ac.uk).

T. X. Vu is with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg, (e-mail: thang.vu@uni.lu)

H. Shin is with the Department of Electronics and Information Convergence Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Republic of Korea (e-mail: hshin@khu.ac.kr).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K., and also with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, South Korea (e-mail: tduong@mun.ca).

The work of D. V. Huynh and T. Q. Duong was supported in part by by the Canada Excellence Research Chair (CERC) Programm, project number CERC-2022-00109. The work of S. Cotton was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through the EPSRC Hub on All Spectrum Connectivity (EP/X040569/1). The work of T. X. Vu was funded in part by the Luxembourg National Research Fund (FNR), grant referenceFNR/C22/IS/17220888/RUTINE. The work of O. A. Dobre was supported in part through the Canada Research Chairs Program, project number CRC-2022-00187. The work of H. Shin was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2022R1A4A3033401).

gap between theoretical evaluation and real-world deployment, where ES capacity is restricted, and the environment is highly unpredictable [20].

Recently, integrated sensing and communication (ISAC) has emerged as a promising technology for future wireless networks, wherein base stations (BS) undertake dual functions encompassing sensing and communication. The ISAC technology has potential to enable new applications across many different domains. This includes enhancing localisation and tracking, drone monitoring, smart home and in-cabin sensing, vehicle-to-everything (V2X) communication, smart manufacturing, industrial Internet-of-Things (IoT), remote sensing, and geoscience, as well as human-computer interaction [21]. However, there exist many open challenges within ISAC systems. Among these are fundamental theories pertaining to radar sensing and wireless communication within the integrated ISAC framework, optimal physical-layer system design for ISAC systems, and optimal cross-layer design for performance optimisation [22], [23]. Consequently, numerous research groups are focusing on studying this topic to enable real-world applications built upon the concept of ISAC [24].

A. Related Works

The convergence of MEC and advanced wireless communication technologies has recently garnered significant attention from the research community. Publications in this domain primarily focus on resource allocation solutions for the joint computing and communication problem [3], [15], [25]–[27]. Specifically, a distributed computation and communication resource management solution was introduced in [25] to address a fairness-aware latency minimisation problem in digital twin-aided MEC systems. In [3], the proposed iterative optimisation solution effectively tackled a sum-utility maximisation problem in computation-intensive systems by optimising computation offloading and service caching. Furthermore, an alternating optimisation algorithm was applied to solve the latency minimisation problem in [15], where various computing and communication variables such as edge selection, power allocation, task offloading portion, and computing capacity allocation were jointly optimised. A learning-based approach was introduced in [26] with edge-assisted spectrum sharing for freshness-aware industrial wireless networks. Additionally, [27] presented a contextual clustering of bandits-based online vehicular task offloading to minimise the expectation of total offloading energy consumption. However, there are still open challenges to explore further in this area, such as adaptive service placement for task-oriented MEC systems, cooperation between MEC and advanced wireless communication technologies like URLLC, and ISAC for future industrial IoT applications.

In the realm of ISAC-assisted MEC, recent publications mainly concentrate on the joint optimal design of computing and communication resources, aiming either at maximising energy efficiency or minimising energy consumption [28]–[30]. Specifically, a joint optimisation problem involving beamforming for radar sensing and task offloading has been addressed to maximise the energy efficiency of ISAC devices [28]. Moreover, the intelligent reflecting surface (IRS) technology has been utilised to support ISAC-based MEC [29]. Here an alternating optimisation algorithm has been proposed to address an optimisation problem concerning the joint allocation of computing and communication resources aimed at minimising energy consumption. More recently, another energy consumption minimisation problem, formulated based on MEC-assisted ISAC with short-packet transmissions, has been tackled [30]. In [30], a hierarchical optimisation procedure is introduced to address various variables, including beamforming designs, computing resource allocation, and transmission duration. In summary, recent efforts in the research of ISAC-assisted MEC have mainly focused on resolving optimisation problems related to joint computing and communication resource allocation. While advanced wireless communication technologies such as IRS and short-packet transmissions have received some attention, adaptive solutions with service placement in MEC remain open issues to be addressed. Further, how these advanced communications technologies can work in unison with MEC to improve the overall network performance.

B. Motivation and Contributions

As highlighted in the aforementioned studies, the convergence of MEC and advanced communication technologies such as URLLC and ISAC presents both vast opportunities and daunting challenges for the research community. To fully harness the potential of these emerging technologies, it is imperative to develop efficient optimisation solutions for addressing the joint sensing, computing, and communication problems. However, despite the considerable attention garnered by recent studies in MEC, the focus has predominantly been on the computation offloading problem, overlooking the crucial aspect of service placement for adaptive service deployment [3], [15], [25]–[27]. Furthermore, the integration of ISAC technology into MEC-based systems, particularly in contexts with stringent requirements for reliability and critical communications, remains an open challenge that demands concerted efforts to make significant contributions [24], [30]. It is evident that addressing these gaps in research is essential for unlocking the full potential of MEC and advanced communication technologies, thus paving the way for transformative advancements in various domains. Jointly designing ISAC with URLLC and edge computing significantly enhances network reliability and reduces latency, crucial for real-time applications like industrial automation. This integration allows for real-time processing and decision-making at the network edge, minimising the need for data transmission to central servers. Additionally, it optimises resource use by leveraging ISAC for efficient spectrum utilisation and edge computing for local data processing, resulting in a more responsive and efficient network.

In this paper, we propose an adaptive optimisation solution for joint service placement, task offloading, and bandwidth allocation in ISAC-based MEC under URLLC-based transmissions. Specifically, we consider the MEC architecture where base stations perform the dual function of sensing and communication to facilitate task offloading with stringent

URLLC requirements. Notably, we introduce a novel cost metric to minimise not only the number of installed services in ESs but also the total latency of user equipment (UEs). The formulated optimisation takes into account service placement decisions, task offloading portions, and bandwidth allocation while considering dynamic timescales to deal with uncertainties in practical deployments. The main contributions of this work can be summarised as follows:

- We initially formulate an optimisation problem aimed at minimising both the number of installed services in ESs and the total latency experienced by UEs. This problem considers various variables within the ISAC-based service-oriented system, including service placement strategies, task offloading proportions, bandwidth allocation under latency requirements, energy budgets of UEs, and computing capacity budgets of ESs.
- Subsequently, we propose an alternating optimisation algorithm to address this challenging problem. Effective inner approximations are appropriately utilised to handle non-convex functions during the solution development process. Additionally, we introduce a Sequential-Fixing (SF) algorithm as a near-optimal approach for solving the mixed-integer non-linear programming (MINLP) problem associated with service placement optimisation.
- Finally, we conduct extensive simulations to validate the effectiveness of our proposed solutions. The numerical results not only demonstrate the superiority of our approach in optimising the cost metric and minimising latency but also illustrate the impact of various parameters on system performance.

C. Paper Structure and Notations

The paper is structured as follows. It begins with an introduction in the first section, providing an overview of the topics under examination, recent publications in the research community, and the contributions made by this paper. Following this, Section II outlines the system model and problem formulation. This section explains the service placement model, task-oriented MEC model, energy consumption model, URLLC-based transmission model, and radar sensing model, concluding with the presentation of the formulated optimization problem. Subsequently, Section III discusses the development of the proposed solutions. Finally, the paper presents simulation results and discussions before concluding with key highlights.

Notation: In this paper, lowercase letters represent numbers, while matrices and vectors are denoted by bold uppercase and lowercase letters, respectively. We adopt the notation $x \sim \mathcal{CN}(.,.)$ to indicate that x follows a complex circularly symmetric Gaussian distribution. The symbol $|\cdot|$ denotes the Euclidean norm of a vector, and $\mathbb{C}$ signifies the set of complex numbers. Additionally, we use the notation $x_{mk}[t_\ell]$ to denote a variable x associated with the m-th user (UE) the k-th edge server (ES) at the long-term time-frame ℓ.

II. System Model and Problem Formulation

We consider a service-oriented MEC model, as illustrated in Fig. 1. In particular, there are M UEs connected wirelessly

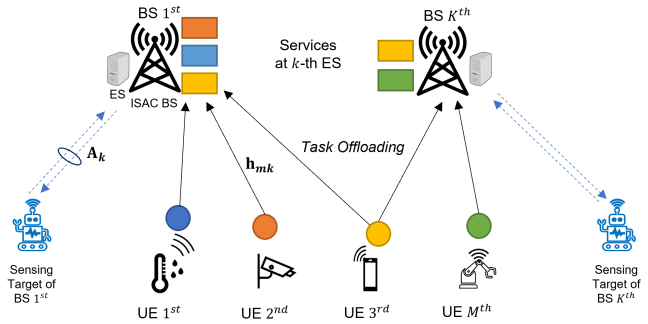

Fig. 1: An illustration of the service-oriented in ISAC-enabled MEC. UEs can partially offload computation tasks to multiple ESs, where the requested services are available at the ESs.

to K ESs via URLLC-based links to guarantee stringent requirements on latency and reliability. Each ES is associated with an L-antenna BS. All UEs are equipped with a single antenna. In our system model, the BS acts as an integrated sensing and communications BS (ISAC-BS) and is able to perform the dual-function of sensing and communications. We assume that there is one sensing target in the coverage area of each BS (e.g., movable robots, autonomous vehicles, etc.).

A. Service Placement Model

The system operates in two discrete time-frames, including long-term frames $\ell \in [1, 2, ..., L]$ for service placement optimisation and short-term time-slots $t \in [1, 2, ..., T]$ for task offloading and resource allocation optimisation. The duration of each long-term frame is Δ_l and each can be divided into T_l short-term time-slots; thus we have $\Delta_l = T_l \delta_t$, where δ_t is the duration of each short-term time-slot. The duration of the long-term time-frames is adaptively adjusted to deal with uncertainties in operation. The optimisation of service placement is only executed when there are new requested services in the system or the latency requirement of UEs is unsatisfied. The proposed approach is capable of reducing the processing cost as well as improving the flexibility in practical implementation.

Each ES can install a finite number of services to execute the offloaded tasks from UEs. The installed services at ES k are denoted by $\mathbf{s}_k[\ell] \triangleq [s_{k1}[\ell], s_{k2}[\ell], ..., s_{kN}[\ell]]$, where N is the total number of services and $s_{kn}[\ell] \in \{0, 1\}, \forall k, n$ indicates the n-th service is installed at ES k (*i.e.* $s_{kn}[\ell] = 1$) or not (*i.e.* $s_{kn}[\ell] = 0$) at long-term frame ℓ. Due to the limited computation capacity of ESs, we consider a partial number of services installed in each ES, yielding $\sum_{n=1}^{N} s_{kn}[\ell] \leq N_k, \forall k$ with $N_k \leq N, \forall k$.

B. Service-Oriented Mobile Edge Computing Model

Let $J_m[t_\ell] \triangleq (T_m^{\max}[t_\ell], D_m[t_\ell], C_m[t_\ell], G_m[t_\ell])$ be the computational task offloaded from the m-th UE, in which $T_m^{\max}[t_\ell]$ is the latency requirement (seconds), $D_m[t_\ell]$ is the task size (bits), $C_m[t_\ell]$ is the required CPU cycles to execute the task (cycles), and $G_m[t_\ell]$ indicates the type of service to execute this task. The computational tasks are only executed

when they are offloaded to correct services in ESs, following $\sum_{k=1}^{K} s_{kG_m}[\ell] \geq 1, \forall k$.

Let us denote the processing rate of the m-th UE and the k-th ES by $f_m^{\mathsf{ue}}[t_\ell]$ and $f_k^{\mathsf{es}}[t_\ell]$, respectively. In this paper, we consider partial task offloading to execute computational tasks [13], [15]. Let $\boldsymbol{\alpha}[t_\ell] \triangleq \{\alpha_m[t_\ell]\}_{\forall m} \mid 0 \leq \alpha_m[t_\ell] \leq 1, \forall m$ be the portion of the task executed locally at the m-th UE at frame t_ℓ. Then, the local processing latency of the m-th UE is given by

$$T_m^{\mathsf{ue}}[t_\ell] = \frac{\alpha_m[t_\ell] C_m[t_\ell]}{f_m^{\mathsf{ue}}[t_\ell]}, \forall m. \tag{1}$$

By defining $\boldsymbol{\beta}[t_\ell] \triangleq \{\beta_{mk}[t_\ell]\}_{\forall m,k} \mid 0 \leq \beta_{mk}[t_\ell] \leq 1, \forall m, k$ as the portion of the task offloaded from the m-th UE to the k-th ES, the processing latency for executing the offloaded task at the k-th ES can be modelled as

$$T_{mk}^{\mathsf{es}}[t_\ell] = \frac{s_{kG_m}[t_\ell] \beta_{mk}[t_\ell] C_m[t_\ell]}{f_{mk}^{\mathsf{es}}[t_\ell]}, \forall m, k. \tag{2}$$

It is noted that the tasks offloaded from the m-th UE is only executed at ESs which have already installed the service G_m at the time-frame t_ℓ.

C. Wireless Transmission Model

In this paper, we adopt the frequency division multiple access (FDMA) protocol for the wireless transmission over the total system bandwidth, B with the noise density N_0. The portion of bandwidth allocated to the m-th UE by the k-th BS is $b_{mk}[t_\ell]$, satisfying $\sum_{m=1}^{M} \sum_{k=1}^{K} b_{mk}[t_\ell] \leq 1$. The channel vector between the k-th BS and the m-th UE is denoted by $\mathbf{h}_{mk}[t_\ell] = \sqrt{g_{mk}[t_\ell]} \bar{\mathbf{h}}_{mk}[t_\ell]$, where $g_{mk}[t_\ell]$ is the large-scale channel coefficient including the path-loss and shadowing, and $\bar{\mathbf{h}}_{mk}[t_\ell]$ represents the small-scale fading following the Rayleigh fading model as $\bar{\mathbf{h}}_{mk}[t_\ell] \sim \mathcal{CN}(0, \mathbf{I}_L)$. Similarly, the sensing channel between the k-th BS and its sensing target is denoted by $\mathbf{d}_k \in \mathbb{C}^{L \times 1}$.

1) Radar Sensing Model: The ISAC-BS simultaneously conducts radar sensing and receives offloaded tasks from IoT devices on the same spectrum channel. The joint radar-communications system includes an active, mono-static radar and supports multi-user task offloading transmissions. At the k-th BS, the received signal $\mathbf{y}_k$ comprises the radar sensing signal $\mathbf{y}_k^{\text{sen}}$ and the offloaded transmission signal $\mathbf{y}_k^{\text{com}}$, represented as:

$$\mathbf{y}_k = \mathbf{y}_k^{\text{sen}} + \mathbf{y}_k^{\text{com}} + \mathbf{n}_k, \tag{3}$$

where $\mathbf{n}_k$ is the additive white Gaussian noise (AWGN) with zero mean and variance $b_{mk} B N_0$.

We begin by analysing $\mathbf{y}_k^{\text{sen}}$. The radar signal transmitted by the BS is denoted as x_k^{sen}. Matrix $\mathbf{A}_k \in \mathbb{C}^{L \times L}$ represents the target response matrix for the radar. Following the assumptions in [31], we consider that we are tracking the target with some prior knowledge of its range. To process the radar signal, a predicted radar return is generated using the predicted target range, which is then subtracted from the received signal to derive a suppressed radar return signal $\tilde{x}_k^{\text{sen}}$ [31]. Consequently, the radar sensing signal $\mathbf{y}_k^{\text{sen}}$ is expressed as:

$$\mathbf{y}_k^{\text{sen}} = \mathbf{A}_k \mathbf{w}_k \tilde{x}_k^{\text{sen}}, \tag{4}$$

where $\mathbf{w}_k \in \mathbb{C}^{L \times 1}$ is the beamforming vector of the radar signal. Following that, we have $\mathbf{A}_k = \mathbf{d}_k \mathbf{d}_k^H$ and $\mathbf{w}_k = \mathbf{d}_k^H / \|\mathbf{d}_k\|$, employed by the maximum-ratio transmission (MRT).

For task offloading communications, $\mathbf{y}_k^{\text{com}}$ at the k-th BS is modelled as

$$\mathbf{y}_k^{\text{com}} = \sqrt{p_m} \mathbf{h}_{mk} x_{mk}^{\text{com}}. \tag{5}$$

Consequently, the received signal $\mathbf{y}_k$ at the k-th ISAC-BS is modelled as follows

$$\mathbf{y}_k = \underbrace{\sqrt{p_m} \mathbf{h}_{mk} x_{mk}^{\text{com}}}_{\text{the desired signal}} + \underbrace{\mathbf{A}_k \mathbf{w}_k \tilde{x}_k^{\text{sen}}}_{\text{the interference of radar signal}} + \mathbf{n}_k. \tag{6}$$

Under FDMA, the signal-to-interference-plus-noise ratio (SINR) ratio of the signal received from the m-th UE at the k-th BS is given by

$$\begin{aligned} \gamma_{mk}(b_{mk}) &= \frac{p_m \|\mathbf{h}_{mk}\|^2}{\rho^2 (b_{mk} B)^2 \sigma_{\text{pre}}^2 \|\mathbf{A}_k \mathbf{w}_k\|^2 + b_{mk} B N_0} \\ &= \frac{p_m \|\mathbf{h}_{mk}\|^2}{\Phi(b_{mk})}, \end{aligned} \tag{7}$$

where p_m is the uplink transmission power of the m-th UE, and $\rho^2 (b_{mk} B)^2 \sigma_{\text{pre}}^2$ is the variance of $\tilde{\mathbf{x}}_k^{\text{sen}}$ [31].

2) URLLC-based transmission model: As a result, the URLLC-based uplink transmission rate (bits/s) for task offloading from the m-th UE to the k-th BS can be expressed as [16], [32]:

$$\begin{aligned} R_{mk}(b_{mk}[t_\ell]) = \frac{B}{\ln 2} \Big[& b_{mk}[t_\ell] \ln\big(1 + \gamma_{mk}(b_{mk}[t_\ell])\big) \\ & - \sqrt{\frac{b_{mk}[t_\ell] V_{mk}(b_{mk}[t_\ell])}{\phi B}} Q^{-1}(\epsilon_{mk}[t_\ell]) \Big], \end{aligned} \tag{8}$$

where ϕ is the transmission time interval, $V_{mk}(b_{mk}[t_\ell]) = 1 - [1 + \gamma_{mk}(b_{mk}[t_\ell])]^{-2}$ is the channel dispersion function, $Q^{-1}(.)$ is the inversion function of $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp^{\left(\frac{-u^2}{2}\right)} du$, and ϵ_{mk} is the decoding error probability.

As a result, the transmission latency from the m-th UE to the k-th BS for task offloading can be expressed as

$$T_m^{\mathsf{co}}[t_\ell] = \frac{s_{kG_m}[\ell] \beta_{mk}[t_\ell] D_m[t_\ell]}{R_{mk}(b_{mk}[t_\ell])}. \tag{9}$$

D. Latency and Energy Consumption Model

1) Latency Model: The end-to-end (e2e) latency incurred in the computation of the task for the m-th UE consists of the local processing latency, the wireless transmission latency and the edge processing latency, *i.e.*

$$T_m^{\mathsf{e2e}}[t_\ell] = T_m^{\mathsf{ue}}[t_\ell] + \max_{\forall k} \{T_m^{\mathsf{co}}[t_\ell]\} + \max\{T_{mk}^{\mathsf{es}}[t_\ell]\}. \tag{10}$$

Since each UE can offload tasks to multiple ESs for execution if these ESs have the indicated services, we use the maximum operator $(\max(.))$ to calculate the worst-case latency of the wireless transmission and edge processing latency. We note that in edge computing, the returned response from ESs to UEs are typically small (*e.g.* control packets and the results

of computations) whereas the transmission ability of APs are significantly greater than UEs; therefore, the downlink transmission latency is not considered in this work [33].

2) Energy Consumption Model of UEs: The energy consumption of the m-th UE includes the energy for the local processing (E_m^{comp}) and wireless transmission (E_m^{comm}), and is given as

$$\begin{aligned} E_m[t_\ell] &= E_m^{\mathsf{comp}} + E_m^{\mathsf{comm}} \\ &= \frac{\theta_m}{2}\alpha_m[t_\ell]C_m[t_\ell](f_m[t_\ell])^2 \\ &\quad + p_m \sum_{k\in\mathcal{K}} \frac{s_{kG_m}[\ell]\beta_{mk}[t_\ell]D_m[t_\ell]}{R_{mk}(b_{mk}[t_\ell])} \end{aligned} \tag{11}$$

where $\theta_m/2$ is the term which accounts for the computation energy consumption of the m-th UE (Watt.s^3/cycle^3).

3) Cost Metric: In this paper, we consider a novel cost metric serving as an objective function of the optimisation problem that aims to minimise the total e2e latency of UEs as well as number of installed services at each ES. To do that, the cost metric η_k is modelled as follows

$$\begin{aligned} \eta_k(\mathbf{s}[\ell],&\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]) = w_k^{\omega}\sum_{n\in\mathcal{N}} s_{kn}[\ell] \\ &+ w_k^{t}\sum_{m\in\mathcal{M}} T_m^{\mathsf{e2e}}(\mathbf{s}[\ell],\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]), \forall k \end{aligned} \tag{12}$$

where w_k^{ω} and w_k^{t} are weights of the number of services and the total latency, respectively. In simulations, the weights are appropriately adjusted to balance the objective components and improve the optimal solutions while executing the proposed algorithm. The weighting parameters are chosen based on the simulation scenarios, e.g., maximum number of services in the system, the total latency of all UEs. The main goal of setting the weights is to create the balance between total number of installed services and the total latency of UEs in the expression of cost metric function.

E. Optimisation Problem Formulation

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{s},\mathbf{b}} \sum_{k=1}^{K}\eta_k(\mathbf{s}[\ell],\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]), \forall t_\ell \tag{13a}$$

$$\text{s.t. } T_m^{\mathsf{e2e}}(\mathbf{s}[\ell],\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]) \le T_m^{\max}, \forall m, t_\ell \tag{13b}$$

$$\alpha_m[t_\ell] + \sum_{k\in K} s_{kG_m}[\ell]\beta_{mk}[t_\ell] = 1, \forall m, t_\ell \tag{13c}$$

$$1 \le \sum_{n=1}^{N} s_{kn}[\ell] \le N_k^{\max}, \forall k, \ell \tag{13d}$$

$$\sum_{k=1}^{K}\sum_{m=1}^{M} b_{mk}[t_\ell] \le 1, \forall k, t_\ell \tag{13e}$$

$$R_{mk}\left(\mathbf{b}[t_\ell]\right) \ge R_{\min}, \forall m, k, t_\ell \tag{13f}$$

$$E_m\left(\mathbf{s}[\ell],\alpha_m[t_\ell],\boldsymbol{\beta}_k[t_\ell]\right) \le E_m^{\max}, \forall m, t_\ell \tag{13g}$$

$$\sum_{m\in\mathcal{M}} s_{kG_m}[\ell]\beta_{mk}[t_\ell]f_{mk}^{\mathsf{es}} \le F_{\max}^{\mathsf{es}}, \forall k, t_\ell \tag{13h}$$

$$\mathbf{s}[\ell]\in\mathscr{S}, \boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell]\in\mathscr{D}, \mathbf{b}[t_\ell]\in\mathscr{B}. \tag{13i}$$

In this paper, we aim to minimise both number of installed services at ES and the total e2e latency of UEs subject to the latency requirements of the computational tasks and the resource budget of the system. The considered optimisation problem is therefore formulated as (13).

In problem (13), constraints (13b) and (13c) indicate the maximum latency requirement and the offloading policies of each computational task, respectively. Constraint (13d) is used for the service placement decision in the system model. The system budget of bandwidth allocation is guaranteed by constraint (13e). The quality-of-service (QoS) requirements for the wireless transmission rate and the energy budget of the UEs are given by constraints (13f) and (13g), respectively. Finally, constraint (13i) incorporates the feasible sets of the optimisation variables defined as follows, $\mathscr{S} \triangleq \{s_{kn} \mid s_{kn} \in \{0,1\}\}, \forall k, n$, $\mathscr{D} \triangleq \{\alpha_m, \beta_{mk} \mid 0 \le \alpha_m \le 1, 0 \le \beta_{mk} \le 1, \forall m, k\ \}$, and $\mathscr{B} \triangleq \{b_{mk} \mid 0 \le b_{mk} \le 1, \forall m, k.\}$

III. Proposed Solutions

As we can see from problem (13), this is a mixed-integer (binary) non-convex optimisation problem which includes strong coupling binary and continuous variables (*e.g.*, (13a), (13c), (13g), (13h)) as well as highly complicated non-convex constraints (*e.g.*, (13b), (13f), (13g)). These challenges make the problem computationally intractable and inefficient to solve directly. Based on the structure of the underlying problem and the properties in practical implementation, we therefore propose a two-timescale optimisation solution which includes the joint task offloading and bandwidth allocation at fixed short-term time-frames and the service placement optimisation with dynamic long-term time-frames to obtain the optimal solutions. The solution is clearly developed in the following subsections.

A. Short-term Joint Task Offloading and Bandwidth Allocation Optimisation

In this sub-problem, we solve for the optimal values of task offloading and bandwidth allocation $\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]$ at the long-term time-frame ℓ with given $\mathbf{s}[\ell]$. The sub-problem is expressed as follows

$$\mathsf{SP\text{-}1}: \underset{\substack{\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\\ \mathbf{b}[t_\ell]|\mathbf{s}[\ell]}}{\text{minimise}} \sum_{k=1}^{K}\eta_k(\mathbf{s}[\ell],\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[\ell_{+1}],\mathbf{b}[t_\ell]), \forall t_\ell, \tag{14a}$$

$$\text{s.t.(13b), (13c), (13e), (13f), (13g), (13h), (13i).} \tag{14b}$$

We solve problem (14) by applying the successive convex approximation (SCA) method. To do that, we have to convexify all non-convex constraints in (14), including (13b), (13f), (13g). These constraints include the wireless transmission rate $R_{mk}(b_{mk}[t_\ell])$ in (8). Therefore, we start with the convexity of constraint (13f).

Convexity of (13f): Following [12], [16], for a sufficiently high S, we have $V_{mk}(b_{mk}[t_\ell]) \approx 1$. Then, we can rewrite $R_{mk}(b_{mk}[t_\ell])$ as

$$R_{mk}(b_{mk}[t_\ell]) = \frac{B}{\ln 2}\Big[G_{mk}(b_{mk}[t_\ell]) - W_{mk}(b_{mk}[t_\ell])\Big] \tag{15}$$

where $G_{mk}(b_{mk}[t_\ell]) = b_{mk}[t_\ell]\ln\big(1+\gamma_{mk}(b_{mk})[t_\ell]\big)\big)$ and $W_{mk}(b_{mk}[t_\ell]) = Q^{-1}(\epsilon_{mk})\sqrt{b_{mk}[t_\ell]}/\sqrt{\phi B}$.

Following inequality [11, eq. 73], given $x>0, y>0$, and $\bar{x},\bar{y}$ are the feasible points of (x,y), we have

$$x\ln(1+\frac{a}{y}) \geq 2\bar{x}\ln(1+\frac{a}{\bar{y}}) + \frac{\bar{x}a}{a+\bar{y}}(1-\frac{y}{\bar{y}}) - \frac{\ln(1+a/\bar{y})}{x}\bar{x}^2. \tag{16}$$

By letting $x = b_{mk}[t_\ell]$, $\bar{x} = b^{(i)}_{mk}[t_\ell]$, $y = \Phi(b_{mk}[t_\ell])$, $\bar{y} = \Phi(b^{(i)}_{mk}[t_\ell])$, and $a = p_m\|\mathbf{h}_{mk}[t_\ell]\|^2$ we can approximate $G_{mk}(b_{mk}[t_\ell])$ as follows

$$\begin{aligned} G_{mk}(b_{mk}[t_\ell]) &\geq 2b^{(i)}_{mk}[t_\ell]\ln\Big(1+\frac{p_m\|\mathbf{h}_{mk}[t_\ell]\|^2}{\Phi(b^{(i)}_{mk}[t_\ell])}\Big) \\ &+ \frac{b^{(i)}_{mk}[t_\ell]p_m\|\mathbf{h}_{mk}[t_\ell]\|^2}{p_m\|\mathbf{h}_{mk}[t_\ell]\|^2+\Phi(b^{(i)}_{mk}[t_\ell])}\left(1-\frac{\Phi(b_{mk}[t_\ell])}{\Phi(b^{(i)}_{mk}[t_\ell])}\right) \\ &- \frac{\ln(1+p_m\|\mathbf{h}_{mk}[t_\ell]\|^2/\Phi(b^{(i)}_{mk}[t_\ell]))}{b_{mk}[t_\ell]}(b^{(i)}_{mk}[t_\ell])^2 \\ &\triangleq \mathcal{G}^{(i)}_{mk}(b_{mk}[t_\ell]). \end{aligned} \tag{17}$$

By applying the following equality

$$\sqrt{x} \leq \frac{\sqrt{\bar{x}}}{2} + \frac{x}{2\sqrt{\bar{x}}}, \tag{18}$$

which $x = b_{mk}[t_\ell] > 0$ and $\bar{x} = b^{(i)}_{mk}[t_\ell] > 0$, we can innerly approximate $W_{mk}(b_{mk}[t_\ell])$ as follows

$$\begin{aligned} W_{mk}(b_{mk}[t_\ell]) &\leq \kappa_{mk}\left(\frac{\sqrt{b^{(i)}_{mk}[t_\ell]}}{2} + \frac{b_{mk}[t_\ell]}{2\sqrt{b^{(i)}_{mk}[t_\ell]}}\right) \\ &\triangleq \mathcal{W}^{(i)}_{mk}(b_{mk})[t_\ell]), \end{aligned} \tag{19}$$

where $\kappa_{mk} = Q^{-1}(\epsilon_{mk})/\sqrt{\phi B}$.

Consequently, the transmission rate $R_{mk}(b_{mk}[t_\ell])$ can be innerly convexified as follows

$$\begin{aligned} R_{mk}(b_{mk}[t_\ell]) &\geq \frac{B}{\ln 2}[\mathcal{G}^{(i)}_{mk}(b_{mk}[t_\ell]) - \mathcal{W}^{(i)}_{mk}(b_{mk})[t_\ell])] \\ &\triangleq R^{(i)}_{mk}(b_{mk}[t_\ell]). \end{aligned} \tag{20}$$

As a result, (13f) can be expressed as

$$R^{(i)}_{mk}(b_{mk}[t_\ell]) \geq R_{\min}, \forall m,k,t_\ell, \tag{21}$$

which is now a convex constraint.

Convexity of (13g): By introducing $\mathbf{r} \triangleq \{r_{mk}[t_\ell]\}_{\forall m,k}$ with $r_{mk}[t_\ell] \geq 1/R^{(i)}_{mk}(b_{mk}[t_\ell]), \forall m,k$, we can equivalently express (13g) as (22a) and (22b):

$$r_{mk}[t_\ell] \geq 1/R^{(i)}_{mk}(b_{mk}[t_\ell]), \tag{22a}$$

$$\begin{aligned} &p_m\sum_{k\in\mathcal{K}} s_{kG_m}[\ell]D_m[t_\ell]\beta_{mk}[t_\ell]r_{mk}[t_\ell] \\ &+ \frac{\theta_m}{2}\alpha_m[t_\ell]C_m[t_\ell](f_m[t_\ell])^2 \leq E^{\max}_m. \end{aligned} \tag{22b}$$

However, constraint (22b) is still non-convex. We apply the following inequality

$$xy \leq \frac{1}{2}\left(\frac{\bar{y}}{\bar{x}}x^2 + \frac{\bar{x}}{\bar{y}}y^2\right) \tag{23}$$

to convexify (22b). Given $x=\beta_{mk}[t_\ell]>0$, $\bar{x}=\beta^{(i)}_{mk}[t_\ell]>0$ $y=r_{mk}[t_\ell]>0$, $\bar{y}=r^{(i)}_{mk}[t_\ell]>0$, (22b) is approximated as

$$\begin{aligned} &p_m\sum_{k\in\mathcal{K}} s_{kG_m}[\ell]D_m[t_\ell] \\ &\times\frac{1}{2}\left(\frac{r^{(i)}_{mk}[t_\ell]}{\beta^{(i)}_{mk}[t_\ell]}\big(\beta_{mk}[t_\ell]\big)^2 + \frac{\beta^{(i)}_{mk}[t_\ell]}{r^{(i)}_{mk}[t_\ell]}\big(r_{mk}[t_\ell]\big)^2\right) \\ &+\frac{\theta_m}{2}\alpha_m[t_\ell]C_m[t_\ell](f_m[t_\ell])^2 \leq E^{\max}_m, \end{aligned} \tag{24}$$

which is a convex constraint.

Convexity of (13b): By using $\mathbf{r}$ defined in (22a), we have

$$\begin{aligned} T^{\mathsf{e2e}}_m[t_{\ell+1}] \leq{}& \frac{\alpha_m[t_{\ell+1}]C_m[t_{\ell+1}]}{f^{\mathsf{ue}}_m[t_{\ell+1}]} + \\ &\frac{s_{kG_m}[t_{\ell+1}]\beta_{mk}[t_{\ell+1}]C_m[t_{\ell+1}]}{f^{\mathsf{es}}_{mk}[t_{\ell+1}]} + \\ &p_m\sum_{k\in\mathcal{K}} s_{kG_m}[\ell]D_m\beta_{mk}[t_{\ell+1}]r_{mk}[t_{\ell+1}]. \end{aligned} \tag{25}$$

By applying (23) for (25) with $x=\beta_{mk}[t_{\ell+1}]$, $\bar{x}=\beta^{(i)}_{mk}[t_{\ell+1}]$, $y=r_{mk}[t_{\ell+1}]$, and $\bar{y}=r^{(i)}_{mk}[t_{\ell+1}]$, we have

$$\begin{aligned} T^{\mathsf{e2e}}_m[t_{\ell+1}] \leq{}& \frac{\alpha_m[t_{\ell+1}]C_m[t_{\ell+1}]}{f^{\mathsf{ue}}_m[t_{\ell+1}]} \\ &+ \frac{s_{kG_m}[t_{\ell+1}]\beta_{mk}[t_{\ell+1}]C_m[t_{\ell+1}]}{f^{\mathsf{es}}_{mk}[t_{\ell+1}]} \\ &+ p_m\sum_{k\in\mathcal{K}} s_{kG_m}[\ell]D_m\xi_{mk} \triangleq \mathcal{T}^{(i)}_{mk}, \end{aligned} \tag{26}$$

where $\xi_{mk} = \frac{1}{2}\left[\frac{r^{(i)}_{mk}[t_\ell]}{\beta^{(i)}_{mk}[t_\ell]}\big(\beta_{mk}[t_\ell]\big)^2 + \frac{\beta^{(i)}_{mk}[t_\ell]}{r^{(i)}_{mk}[t_\ell]}\big(r_{mk}[t_\ell]\big)^2\right]$ and $\mathcal{T}^{(i)}_{mk}$ is a convex function.

As a result, we can solve the following convex program to obtain the optimal solution for SP-1 at the i-th iteration:

SP-1-Convex:

$$\underset{\substack{\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\\ \mathbf{b}[t_\ell]|\mathbf{s}[\ell]}}{\text{maximise}} \sum_{k=1}^{K}\Big(w^{\omega}_k\sum_{n\in\mathcal{N}} s_{kn[\ell]} + w^{t}_k\sum_{m\in\mathcal{M}}\mathcal{T}^{(i)}_{mk}\Big), \tag{27a}$$

$$\text{s.t.} \quad \mathcal{T}^{(i)}_{mk} \leq T^{\max}_m, \forall m,k, \tag{27b}$$

$$(13c),(13g),(21),(22a),(24). \tag{27c}$$

In (27), all the convex constraints are linear and/or quadratic, and therefore it can be efficiently solved by the well-known CVX package in the MATLAB environment.

B. Long-term Service Placement Optimisation

In this sub-problem, we find the optimal value of service placement decisions $\mathbf{s}[\ell+1]$ for given $\boldsymbol{\alpha}[t_\ell], \boldsymbol{\beta}[t_\ell], \mathbf{b}[t_\ell]$, which is expressed as

$$\text{SP-2:} \quad \underset{\substack{\mathbf{s}[\ell+1]|\\ \boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[t_\ell],\mathbf{b}[t_\ell]}}{\text{minimise}} \sum_{k=1}^{K}\eta_k(\mathbf{s}[\ell],\boldsymbol{\alpha}[t_\ell],\boldsymbol{\beta}[\ell+1],\mathbf{b}[t_\ell]) \tag{28a}$$

$$\text{s.t.} \quad (13b),(13c),(13d),(13g),(13h),(13i). \tag{28b}$$

The problem (28) is a mixed-integer program. There are several effective approaches to deal with this kind of problem,

such as the relaxation or parameterisation.

Based on the above development, we propose Algorithm 1 to solve (13). The most challenging issue in the implementation of Algorithm 1 is to handle the uncertainties of new services requested and triggering the long-term optimisation when the latency requirement is unmet.

Algorithm 1 : Proposed Algorithm for Solving (13).

1: **Initialisation**: Set $\ell = 1$, $t = 1$, generate the initial feasible points $(\mathbf{s}^{(1)}[1], \boldsymbol{\alpha}^{(1)}[1], \boldsymbol{\beta}^{(1)}[1], \mathbf{b}^{(1)}[1])$, and choose the initial parameters for (13).
2: **while** *(long-term flag is FALSE)* **do**
3: **for** $t = 1, 2, ..., T$ **do**
4: Solve SP-1-Convex (27) for $(\boldsymbol{\alpha}[t_\ell], \boldsymbol{\beta}[t_\ell], \mathbf{b}[t_\ell])$ with given $\mathbf{s}[\ell]$;
5: Set long-term flag TRUE if: (1) new requested services OR (2) latency unsatisfied;
6: **if** *(long-term flag is TRUE)* **then**
7: Solve SP-2 (28) for service placement $\mathbf{s}(\ell + 1)$ with given $(\boldsymbol{\alpha}[t_\ell], \boldsymbol{\beta}[t_\ell], \mathbf{b}[t_\ell])$;
8: Update long-term flag;
9: **end if**
10: Set $t = t + 1$
11: **end for**
12: Set $\ell = \ell + 1$
13: **end while**

Notes on the initialisation and the algorithm complexity: In the initialisation step, we set the offloading portion equally for all UEs, i.e., $\alpha_m = 0.3, \forall m, \beta_{mk} = 0.7/M, \forall m, k$, and the bandwidth is also equally allocated among all UEs, i.e, $b_{mk} = 1/MK$. Regarding the requested service, we assume that the ESs have already installed neccessary services to serve UEs in the first time-frame. To guarantee the success of the initialisation, we implement a function to check if all constraints in the orginal problem (13) are satisfied or not before the algorithm proceeds with the next step.

In relation to the algorithm complexity, the short-term joint task offloading and bandwidth allocation problem (27) has a total of $2MK + M$ scalar variables and $3MK + 3M$ constraints; therefore, the per-iteration complexity for solving it is $\mathcal{O}\big(\sqrt{3MK + 3M}\ (2MK + M)^2\big)$ [34, Sec. 6]. Similarly, the per-iteration of the problem (28) is $\mathcal{O}\big(\sqrt{KN + 3M + 2K}\ (NK)^2\big)$.

C. Near-optimal Design for the Mixed-Integer Service Placement Optimisation Problem

In this subsection, we propose a sequential fixing (SF)-based solution for the MINLP of the service placement problem. This is a heuristic procedure and has polynomial-time complexity, which is an efficient technique to deal with MINLPs [35]. This heuristic method balances the need for computational efficiency with the goal of finding a high-quality solution. The method is highly dependent on the initial relaxation and variable selection strategies, which can significantly influence the quality of the final solution. Additionally, it may require numerous iterations and adjustments, particularly if feasibility issues arise, leading to potentially high computational costs. Regarding applying this method to the paper, the main idea is to fix the values of $s_{km}[\ell]$ sequentially by solving a series of linear program (LP) problems and set at least one binary value for some $s_{km}[\ell]$ during each iteration. For instance, during the first iteration of long-term solving, all binary variables of $\mathbf{s}[\ell]$ are relaxed to satisfy $0 \leq s_{km}[\ell] \leq 1, \forall m, k$ to transform (28) into an LP problem. After solving this LP problem, we obtain a value between 0 and 1 for each variable $s_{km}[\ell]$. The procedure is repeated until each ES reachs the maximum number of installed services (i.e., constraint (13d)). The solving process is summarised as the following Algorithm 2.

Algorithm 2 : The SF-based Algorithm for Solving the MINLP (28) problem.

1: **Initialisation**: Set up and solve the initial relaxed LP problem of (13) with variables $0 \leq s_{km}[\ell] \leq 1, \forall m, k$.
2: Suppose $s_{km}[\ell]$ is the largest value among all the s-variables, fix $s_{km}[\ell] = 1$.
3: **if** (all the s-variables are fixed) **then**
4: Return the obtained solutions.
5: **end if**
6: Reformulate and solve a new relaxed LP problem with the newly fixed s-variables and go to Step 2.

IV. Simulations, Results and Discussions

A. Simulation Settings

In the following simulations, we consider a system model for industrial automation where all ESs and UEs are distributed within an area of 100m $\times$ 100 m. The large-scale fading for the wireless transmission between the m-th UE to the k-th ES is modelled as $g_{mk} = 10^{\mathbf{PL}(\mathrm{d_{mk}})/10}$, where $\mathbf{PL}(d_{mk}) = -35.3 - 37.6 \log_{10} d_{mk}$ [11]. The single-sided noise spectral density is set to -174 dBm/Hz [11]. The URLLC decoding error probability is set to $\epsilon_{mk} = 10^{-6}$ [36]. The other simulation parameters are provided in Table I. All simulations were conducted in MATLAB and the convex programs were solved by the CVX package.

TABLE I: Simulation Parameters.

Parameters	Value
Number of antennas	$L = 8$
Transmission power	$p_m = 23$ dBm
System bandwidth	$B = 10$ MHz
Number of UEs	$M = 8$
Number of ESs	$K = 2$
Maximum number of services	$N = 6$
UE processing rate	$f_m^{\mathsf{ue}} = 1$ GHz
ES processing rate	$f_{mk}^{\mathsf{es}} = 2$ GHz
Input task size	$D_m = 1354$ bytes [37]
Task complexity	$C_m/D_m = [200, 500]$ cycles/byte
Total delay requirement	$T_m^{\max} = 2$ ms
Minimum data rate	$R^{\min} = 1$ Mbps
Maximum energy consumption	$E_m^{\max} = 1$ Joule
Effective capacitance coefficient	$\theta_m = 10^{-27}$ Watt.s^3/cycle^3 [38]
Radar spectral shape parameter	$\rho = 2\pi/12$ [29]
The variance of the predicted radar return	$\sigma_{\mathrm{pre}}^2 = 10^{-14}$ [29]

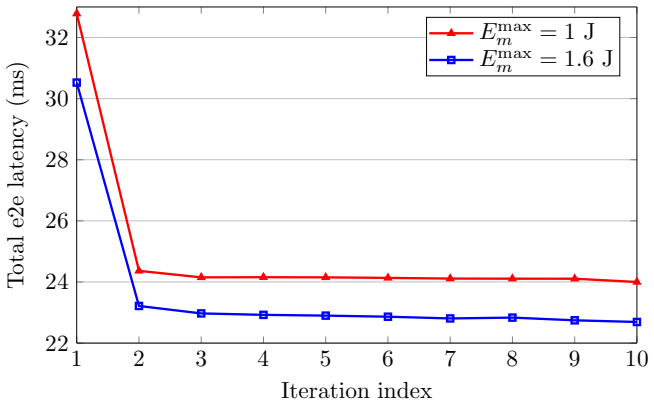

Fig. 2: Convergence pattern of the short-term optimisation for $E_m^{\max} = 1$ J and $E_m^{\max} = 1.6$ J.

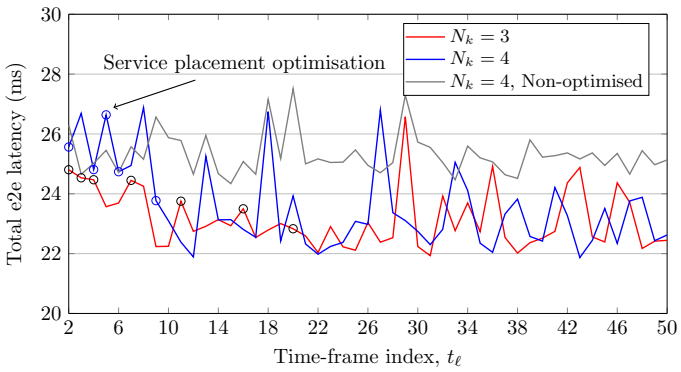

Fig. 3: Impact of the maximum number of services installed in ESs and the service placement optimisation on the obtained latency for $N = 6$ services.

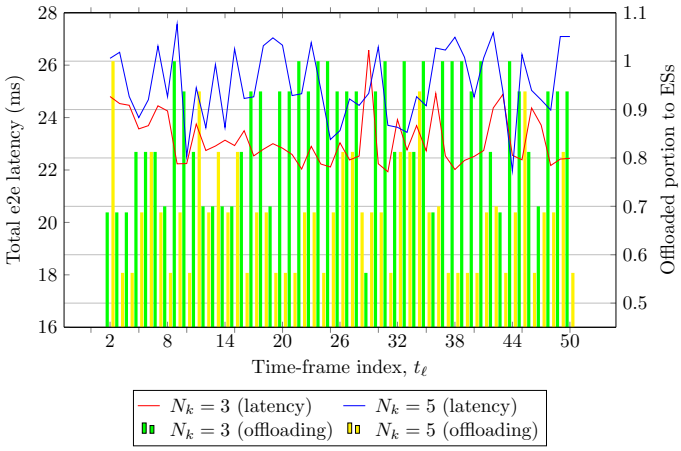

Fig. 4: Impacts of the maximum number of services installed in ESs on the obtained latency and offloading behaviours.

B. Numerical Results and Discussions

In this subsection, we present results of simulations to illustrate the impact of various parameters on system performance, focusing on total latency and the cost metric. We conducted extensive simulations to validate the effectiveness of the proposed solutions and explore factors such as the number of installed services in ESs, optimal service placement, maximum computing capacity of ESs, energy budget of UEs, sensing function in ISAC-based systems, and task complexity of service-oriented systems.

1) Convergence pattern of short-term optimisation: To illustrate the convergence pattern of the proposed iterative algorithm (Algorithm 1), the total e2e latency of UEs obtained after each iteration is recorded throughout the process. Figure 2 clearly exhibits the convergence of the algorithm as it progressively diminishes the total latency. In our implementation, convergence is deemed achieved when the difference in total latency between the current iteration and the previous one is sufficiently small (i.e., $\epsilon = 10^{-3}$) relative to the total latency of the current iteration. As depicted in the graph, the algorithm effectively optimises the total latency to reach optimal values and achieves convergence after 10 iterations. It is noteworthy that the latency experiences a significant decrease after the first iteration and gradually reduces until convergence. This is attributed to the initial values being set uniformly to satisfy all constraints, which are initially far from optimal solutions.

2) Impact of the maximum number of services installed in ESs and the optimal solutions on the latency: To demonstrate the impact of service placement optimisation on the optimal e2e latency, we conducted simulations with different settings of ESs capacity, varying the maximum number of installed services (N_k). Specifically, Figure 3 illustrates the total e2e latency of UEs over a period of 50 time-frames with $N_k = 3, 4$. As evident from the figure, the proposed solution significantly outperforms non-optimised schemes (such as no service placement and equal bandwidth allocation) in minimising latency. Furthermore, when ESs can accommodate a higher number of services, the service placement optimisation is less frequently executed. However, this may lead to a decrease in the overall performance due to the limitation of computing capacity of ESs.

3) Impact of the maximum number of installed services on the latency and offloading behaviours: To further examine the impact of the number of installed services in ESs, we monitored the offloading behaviour of UEs over time with different settings of N_k. Figure 4 depicts the offloaded portion of computational tasks and the total e2e latency over 50 time frames. In particular, the maximum number of services installed in ESs significantly influences the optimal latency by adjusting the offloaded portion of tasks. In Figure 4, the latency obtained in the $N_k = 5$ scenario is higher than that in the $N_k = 3$ scenario due to the computing capacity limitation of ESs (i.e., (13h)). This difference is evident from the bar components of the graph. It is observed that the offloaded portion in the $N_k = 3$ scenario is considerably higher than that in the $N_k = 5$ scenario, confirming the effectiveness of the proposed optimal task offloading solutions under the edge computing capacity budget.

4) Impact of the ES's processing rate and the UE's energy budget: Figure 5 illustrates how the computing capacity of ESs and the UE's energy budget affect e2e latency of UEs. From the chart, it is evident that both the processing rate of ESs and the energy budget of UEs significantly contribute to the system's performance in reducing latency. Specifically, as ESs

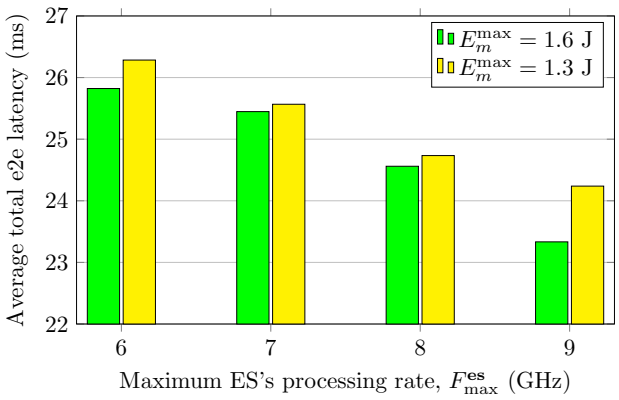

Fig. 5: Impact of the ES's processing rate on the obtained latency for $N = 6, N_k = 3$ services.

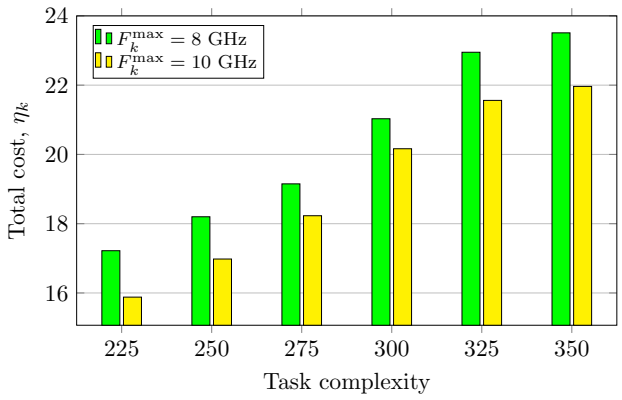

Fig. 6: Impact of the task complexity on the cost metric.

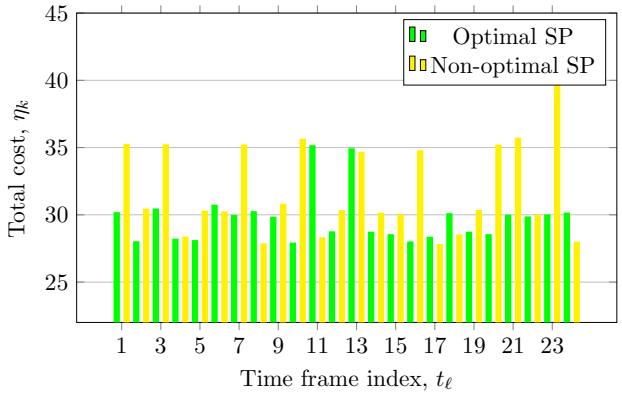

Fig. 7: Impacts of optimal service placement optimisation on the cost metric in the scenarios of $N = 6, N_k = 3$ services.

become more powerful, i.e., with an increased processing rate, the total latency of UEs gradually decreases. This observation confirms that the task offloading solution operates correctly and effectively. Regarding the energy budget of UEs, Figure 5 demonstrates that when the maximum energy consumption of UEs increases, the latency is reduced. These results can be clearly explained by (11) and (13g). When UEs have a larger energy budget, the proportion of local processing increases, resulting in a reduction in transmission latency for offloading tasks to ESs. Consequently, the total e2e latency obtained sustainably decreases. Once again, these results clearly demonstrate the effectiveness of the proposed offloading scheme in minimising the total e2e latency.

5) Impact of the task complexity on the cost metric: To demonstrate the impact of task complexity (cycles/byte) on the total cost considered (i.e., η_k), we conducted simulations with different levels of task complexity and computing budgets of ESs (i.e., $F_k^{\max}$). Figure 6 shows how task complexity affects the obtained cost. Clearly, as the task complexity increases, a greater number of CPU cycles are required to execute a particular byte of the task, the cost gradually rises. This is because the required CPU cycles significantly contribute to the processing latency at both the local level (UEs) and the remote level (ESs), as modelled in (10). Therefore, in order to reduce the cost, strategies such as reducing the task complexity or increasing the computing capacity of ESs can be implemented.

6) Impact of the optimal service placement optimisation on the cost metric: To demonstrate the impact of optimal service placement on the considered cost, we conducted simulations comparing scenarios with optimal service placement and without service placement solutions. As depicted in Figure 7, the optimal solution significantly enhances system performance by consistently minimising the cost. Throughout the simulation, the optimal service placement solution achieves nearly a 50% reduction in cost compared to the non-optimal scheme. This improvement is evident from the formulation of the cost metric in (12), where the cost metric is the weighted summation of total latency and the number of installed services in ESs. Without service placement optimisation, the number of installed services in ESs remains unminimised, and the absence of the best service placement strategy fails to guarantee a reduction in latency. Consequently, this increases the cost value in the considered system.

7) Impacts of the sensing function on the cost metric: To examine the impact of the sensing function of an ISAC-based system on the overall performance, we log the total cost value over the running time with sensing and without sensing scenarios as shown in Fig. 8. As we can see from the figure, the sensing function increases the cost metric by around 25%. This result can be explained by the formulation of the transmission rate (8), where the sensing reduces the transmission rate by increasing the interference in the SINR calculation (i.e., (7)). The result also indicates that it is essential to design an effective optimisation solution for ISAC-based systems to mitigate the latency caused by the interference of sensing signal.

V. Conclusion

We have explored the integrated challenge of sensing, computing, and communication within service-oriented networks. Our investigation centres on a system model with practical relevance for the deployment of ISAC-assisted edge networks. These networks are equipped with dual-functional base stations and support various computation-intensive, time-sensitive services. The optimisation problem addressed in our study not only aims to minimise the number of installed

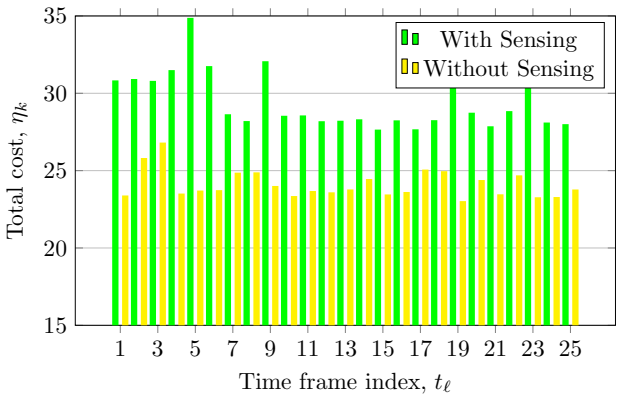

Fig. 8: Impacts of sensing function on the cost metric in the scenarios of $N = 6, N_k = 3$ services, and $F_k^{\max} = 10$ GHz.

services at ESs but also seeks to reduce the e2e latency among UEs. This is achieved while considering constraints such as system budget and unpredictable environmental factors. To tackle this challenge, we have devised an iterative algorithm to determine optimal decisions regarding service placement and resource allocation. Our simulation results underscore the effectiveness of the proposed solution. A promising avenue for future research involves developing machine learning-based solutions for the mixed-integer problem in service placement, especially for medium-to-large network sizes. Additionally, exploring strategies for managing multiple sensing targets in MEC-based systems presents another valuable direction for future studies.

ACKNOWLEDGEMENTS

The work of D. V. Huynh and T. Q. Duong was supported in part by by the Canada Excellence Research Chair (CERC) Programm, project number CERC-2022-00109. The work of S. Cotton was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through the EPSRC Hub on All Spectrum Connectivity (EP/X040569/1). The work of T. X. Vu was funded in part by the Luxembourg National Research Fund (FNR), grant referenceFNR/C22/IS/17220888/RUTINE. The work of O. A. Dobre was supported in part through the Canada Research Chairs Program, project number CRC-2022-00187. The work of H. Shin was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2022R1A4A3033401).

APPENDIX

In this appendix, we provide fundamental inequalities based on the principles of inner approximation [11], [39], which are used to convexify constraints (13f) and (13g). In particular, given the concave function $f(x) = \sqrt{x}$ over $x > 0$, its upper bounding convex function at the point $\bar{x} > 0$ is

$$h(x) \leq f(\bar{x}) + \frac{\partial f(x)}{\partial(x)}\Big|_{x=\bar{x}}(x-\bar{x}) = \frac{\sqrt{\bar{x}}}{2} + \frac{x}{2\sqrt{\bar{x}}}, \quad (29)$$

which has been used in (19).

As the function $g(x,y) = \sqrt{xy}$ is concave on (x,y) with $x > 0, y > 0$, the following inequality holds true for all $x > 0, y > 0, \bar{x} > 0$, and $\bar{y} > 0$:

$$\sqrt{xy} = g(x,y) \leq g(\bar{x},\bar{y}) + \frac{\partial g(x,y)}{\partial(x)}\Big|_{(x,y)=(\bar{x},\bar{y}}(x-\bar{x}) + \frac{\partial g(x,y)}{\partial(y)}\Big|_{(x,y)=(\bar{x},\bar{y}}(y-\bar{y}) = \sqrt{\bar{x}\bar{y}} + \frac{\sqrt{y}}{2\sqrt{x}}(x-\bar{x}) + \frac{\sqrt{x}}{2\sqrt{y}}(y-\bar{y}) = \frac{1}{2}\left(\frac{\sqrt{\bar{x}}}{\sqrt{\bar{y}}}y + \frac{\sqrt{\bar{y}}}{\sqrt{\bar{x}}}x\right), \quad (30)$$

which has been used in (23).

REFERENCES

[1] D. V. Huynh, V.-D. Nguyen, O. A. Dobre, S. R. Khosravirad, and T. Q. Duong, "Adaptive service placement, task offloading and bandwidth allocation in task-oriented urllc edge networks," in *Proc. IEEE Int. Conf. Commun. (ICC'23)*, Rome, Italy, May 28 2023, pp. 5755–5760.

[2] L. Lin, X. Liao, H. Jin, and P. Li, "Computation offloading toward edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1584–1607, Aug. 2019.

[3] S.-W. Ko, S. J. Kim, H. Jung, and S. W. Choi, "Computation offloading and service caching for mobile edge computing under personalized service preference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6568–6583, Aug. 2022.

[4] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2131–2165, Fourth quarter 2021.

[5] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[6] K.-C. Chen, S.-C. Lin, J.-H. Hsiao, C.-H. Liu, A. F. Molisch, and G. P. Fettweis, "Wireless networked multirobot systems in smart factories," *Proc. IEEE*, vol. 109, no. 4, pp. 468–494, Apr. 2021.

[7] 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.

[8] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[9] D. V. Huynh, S. R. Khosravirad, L. D. Nguyen, and T. Q. Duong, "Multiple relay robots-assisted URLLC for industrial automation with deep neural networks," in *Proc. 2021 IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, Dec.7–11 2021.

[10] D. Feng *et al.*, "Toward ultrareliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 94–102, Jun. 2019.

[11] A. A. Nasir, H. D. Tuan, H. Nguyen, M. Debbah, and H. V. Poor, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[12] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 402–415, Jan. 2019.

[13] D. V. Huynh, V.-D. Nguyen, S. Chatzinotas, S. R. Khosravirad, H. V. Poor, and T. Q. Duong, "Joint communication and computation offloading for ultra-reliable and low-latency with multi-tier computing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 2, pp. 521–537, Feb. 2022.

[14] Y. Li, D. V. Huynh, T. Do-Duy, E. Garcia-Palacios, and T. Q. Duong, "Unmanned aerial vehicles-aided edge networks with ultra-reliable low-latency communications: A digital twin approach," *IET Signal Processing*, 2022, accepted.

[15] T. Do-Duy, D. V. 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.

[16] 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.

[17] M. S. Elbamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, "Wireless edge computing with latency and reliability guarantees," *Proc. IEEE*, vol. 107, no. 8, pp. 1717–1737, Aug. 2019.

[18] L. Chen, C. Shen, P. Zhou, and J. Xu, "Collaborative service placement for edge computing in dense small cell networks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 377–390, Feb. 2021.

[19] Z. Ning, P. Dong, X. Wang, S. Wang, X. Hu, S. Guo, T. Qiu, B. Hu, and R. Y. K. Kwok, "Distributed and dynamic service placement in pervasive edge computing networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 6, pp. 1277–1292, Jun. 2021.

[20] H. Ma, Z. Zhou, and X. Chen, "Leveraging the power of prediction: Predictive service placement for latency-sensitive mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6454–6468, Oct. 2020.

[21] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[22] S. Lu, F. Liu, Y. Li, K. Zhang, H. Huang, J. Zou, X. Li, Y. Dong, F. Dong, J. Zhu, Y. Xiong, W. Yuan, Y. Cui, and L. Hanzo, "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet Things J.*, 2024.

[23] K. Meng, Q. Wu, J. Xu, W. Chen, Z. Feng, R. Schober, and A. L. Swindlehurst, "Uav-enabled integrated sensing and communication: Opportunities and challenges," *IEEE Wireless Commun. Mag.*, 2023.

[24] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous iot: Applications, trends, and challenges," *IEEE Netw.*, vol. 35, no. 5, pp. 158–167, Sep. 2021.

[25] D. V. Huynh, V.-D. Nguyen, S. R. Khosravirad, K. Wang, G. K. Karagiannidis, and T. Q. Duong, "Distributed communication and computation resource management for digital twin-aided edge computing with short-packet communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3008–3021, Aug. 2023.

[26] M. Li, C. Chen, H. Wu, X. Guan, and X. S. Shen, "Edge-assisted spectrum sharing for freshness-aware industrial wireless networks: A learning-based approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7737 – 7752, Sep. 2022.

[27] Y. Lin, Y. Zhang, J. Li, F. Shu, and C. Li, "Popularity-aware online task offloading for heterogeneous vehicular edge computing using contextual clustering of bandits," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5422–5433, Apr. 2021.

[28] N. Huang, C. Dou, Y. Wu, L. Qian, and R. Lu, "Energy-efficient integrated sensing and communication: A multi-access edge computing design," *IEEE Wireless Commun. Lett*, vol. 12, no. 12, pp. 2053–2057, Dec. 2023.

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

[30] N. Huang, C. Dou, Y. Wu, L. Qian, B. Lin, H. Zhou, and X. S. Shen, "Mobile edge computing aided integrated sensing and communication with short-packet transmissions," *IEEE Trans. Wireless Commun.*, 2023.

[31] A. R. Chiriyath, B. Paul, G. M. Jacyna, and D. W. Bliss, "Inner bounds on performance of radar and communications co-existence," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 464–474, Jan. 2016.

[32] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[33] R. Lin, Z. Zhou, S. Luo, Y. Xiao, X. Wang, S. Wang, and M. Zukerman, "Distributed optimization for computation offloading in edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8179–8194, Dec. 2020.

[34] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*. Philadelphia: MPS-SIAM Series on Optimi., SIAM, 2001.

[35] Y. T. Hou, Y. Shi, and H. D. Sherali, *Applied Optimization Methods for Wireless Networks*. Cambridge University Press, 2014.

[36] H. Lee and Y.-C. Ko, "Physical layer enhancements for ultra-reliable low-latency communications in 5G new radio systems," *IEEE Commun. Stand. Mag.*, vol. 5, no. 4, pp. 112–122, Dec. 2021.

[37] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.913, 2018, version 15.0.0.

[38] C.-F. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132–4150, Jun. 2019.

[39] V.-D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O.-S. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.