Received 15 December 2021; revised 3 February 2022; accepted 21 February 2022. Date of publication 28 February 2022; date of current version 15 March 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3154536

Incentive Mechanism and Resource Allocation for Edge-Fog Networks Driven by Multi-Dimensional Contract and Game Theories

MARIA DIAMANTI [1] (Graduate Student Member, IEEE), PANAGIOTIS CHARATSARIS [1], EIRINI ELENI TSIROPOULOU [2] (Senior Member, IEEE), AND SYMEON PAPAVASSILIOU [1] (Senior Member, IEEE)

[1]School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Greece

[2]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA

CORRESPONDING AUTHOR: S. PAPAVASSILIOU (e-mail: papavass@mail.ntua.gr)

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members and Researchers and the Procurement of High-Cost Research Equipment Grant" under Project HFRI-FM17-2436.

ABSTRACT The edge computing paradigm has become extremely popular over the past years, as a means of offloading computationally intensive tasks by users of resource and battery-constrained devices. Nevertheless, the edge networks' overexploitation by the ever-increasing number of task-offloading users, gradually leads to their performance degradation. In this paper, we leverage on the different levels of available computing capabilities across the network, and we design an incentive mechanism that aims to shift the selfish users' preference from the edge to the upper fog computing layer, accounting for their level of delay tolerance. To deal with the users' heterogeneity in terms of their applications' multi-dimensional distinctive features (including their delay tolerance/sensitivity), a multi-dimensional contract theory modeling is adopted, according to which the edge server determines the bundles of the users' provided efforts and corresponding offered rewards. In this respect, each user's effort represents the amount of its initially offloaded task at the edge that is allowed to be further forwarded and processed at the fog. Considering that the users-to-edge server offloading is performed under Non-Orthogonal Multiple Access (NOMA), the problem of joint computation task offloading and uplink transmission power allocation is subsequently addressed via a Stackelberg game, where the edge server and the users are treated as leader and followers, respectively. The aim of the game is to minimize the end-to-end network's energy consumption and increase its resource utilization efficiency. The incentive mechanism and resource allocation framework is evaluated via modeling and simulation regarding its operation and efficiency under different scenarios.

INDEX TERMS Computation offloading, edge-fog networks, game theory, incentive mechanism, multi-dimensional contract theory.

I. INTRODUCTION

THE UBIQUITOUS connectivity enabled by the next-generation wireless networks is progressively shaping the frontier of an ambient intelligence era. Striving to reap the benefits brought by the surrounding environment intelligence has provoked the increase of ever computationally intensive user applications. To facilitate the computationally and battery-constrained user devices to meet the corresponding time and energy Quality-of-Service (QoS) requirements, the concept of computation offloading of resource-intensive tasks has become extremely popular. Especially, among the different computing capabilities and options existing within the computing continuum, the Multi-Access Edge Computing (MEC), often implemented

within the Radio Access Network (RAN), has revolutionized the successful completion of low-latency applications [1]. Nevertheless, driven by their appealing properties, the over-exploitation of the edge computing networks will gradually lead to their performance degradation. To alleviate this issue and ameliorate the overall system's resource utilization, a heterogeneous multi-layer computing architecture should be pursued, where different computing entities of various capabilities across the network cooperate with each other.

Indeed, the diversity of the offloaded tasks in terms of their intensity, as well as the heterogeneity of the corresponding user applications' performance requirements regarding their delay (in)sensitivity and power consumption, create a solid ground for the proper utilization of the different computing options across the network. Under this scope, the fog computing, situated anywhere between the network edge and the cloud, appears as the ideal candidate for Internet of Things (IoT) user applications of high processing and storage needs, but of looser completion time and delay constraints [2]. Nevertheless, despite the potential ability of the delay-tolerant tasks to be processed at the fog (or even the cloud) without degrading the QoS, the edge computing's appealing features regarding its proximity to the users along with the users' selfish behavior, may prove to be an impediment in the realization of the envisioned heterogeneous multi-layer computing paradigm [3].

In this paper, we target to exactly address this challenge under a two-layer computing environment, consisting of an edge service layer and a fog service layer that are distinguished from an architectural point of view with respect to the location where their computation power is placed. In particular, in the edge computing model the computational power and intelligence is implemented exactly at the network edge (e.g., local edge), while in the fog computing case this functionality may be offered at different locations between the network edge and the core network connecting to the cloud, thus, exploiting the power of the whole digital continuum (e.g., main edge servers). Within this setting, different users of heterogeneous application performance requirements, wirelessly offload part of their applications' tasks for remote execution at the edge, which in principle from the user perspective prevails against the fog, being just one-hop communication distance away. However, in this paper, leveraging on the extended fog computing capabilities and the wireless communication between the edge and fog servers, we, first, design and propose an incentive mechanism, following the principles of labor economics and multi-dimensional contract theory [4] so that the users exploit the fog computing. Employing the incentive mechanism, the edge server seeks to motivate its offloading users to allow part of their offloaded tasks to be further forwarded and processed at the fog, based on their distinct and heterogeneous applications' characteristics, as a means of improving the resource utilization efficiency across the network, and increasing its overall service capacity, especially under the presence of delay-tolerant services.

Accordingly, we utilize the outcome of the economic interplay between the users-edge-fog layers to tackle the challenging problem of a multi-layer computing environment's resource orchestration. Given the percentage of the initially offloaded tasks at the edge that are allowed to be further forwarded to the fog, the joint computation task offloading and uplink transmission power allocation problem between the users and the edge is addressed, considering the users' transmissions' multiplexing via the Non-Orthogonal Multiple Access (NOMA) technique. Respecting the need for decentralized resource management approaches, we propose a Stackelberg game-theoretic procedure, according to which the edge server (i.e., the leader) derives the optimal amount of task to be actually offloaded at the edge by each user, while at the same time, the users (i.e., the followers) autonomously determine their optimal uplink transmission powers to the edge. The game is played iteratively until convergence, with the overall aim to minimize the end-to-end system's energy overhead, subject to the users' delay-tolerance constraints.

A. RELATED WORK

Several works exist in the literature, dealing with computation task offloading and resource allocation problems in multi-layer computing environments, e.g., [5]–[9]. In [5], the users' full offload of their computation tasks at a primary fog server is assumed and the problem of invoking the assistance of other fog servers or of the cloud is studied to complete the users' tasks within their time constraint. In [6], the multi-user decision problem of their computation tasks' execution either locally, or at the fog or at the cloud is formulated and solved as a potential game, while other works consider similar offloading decision problems at multiple computing layers along with computing resource allocation [7], joint computing resource, uplink transmission power and radio resource allocation [8], or servers' service caching decision [9] problems. On the one hand, none of the existing works accounts for the three-level splitting (i.e., local, edge and fog layers) of the users' computation tasks, while overlooking the economic and market perspective of the computation offloading as a service.

In this paper, though the users are provided with a transparent computing service - meaning that the multiple service layers are viewed as a contiguous computing network - they can still smartly evaluate the emerging tradeoffs between delay tolerance and task intensity and size, which are directly affected by the available computing options. Shifting the selfish users' preference to upper computing layers according to their delay-tolerance levels, calls for the creation and provisioning of appropriate incentives. In this context, a well-established method to deal with the problem of incentives comes from the field of labor economics and contract theory [4]. Under its general form, a contract-theoretic model includes an employer that creates contract bundles tailored to the different employees' personal characteristics in order to motivate them provide back

their efforts. The contracts are designed under "incompleteness of information", meaning that the employees' personal characteristics are initially unknown to the employer, constituting in this way the employees' "private information" or "user types". Among the wide variety of applications of contract theory in wireless communications and networking (e.g., cognitive radio networks [10], Device-to-Device (D2D) communications [11], crowdsourcing [12], resource allocation [13]), some effort has been made in the direction of computation offloading. In [14], the problem of incentivization of potential temporary edge nodes from an edge computing operator is examined under a MEC paradigm. Similar problems are considered in [15], [16] under the concept of vehicular edge/fog computing offered by vehicles to other travelling vehicles or roadside users. Different from the concept of computation offloading, but relevant to the incentivization of delay-tolerant users is the work in [17]. This work suggests that the users capitalize on their delay tolerance and cost sensitivity, and forward their traffic through the available Delay-Tolerant Networks (DTNs) or WiFi networks, in return for reduced service cost.

Focusing on the practical application of contract theory models, most of the existing works in the literature, including the aforementioned ones in [10]–[12], [14]–[17], rely on one-dimensional user types that typically capture each user's level of willingness or ability to participate in the contract. Nevertheless, such an approach appears to be rather restrictive, since in most cases there are more than one distinguishing features for each user that should steer the contract modeling, especially when these features are conflicting. Recently, the problem of multi-dimensional contract theory in terms of the number of user types that characterize each user has been investigated in [18]–[20]. In particular, in [18], the interplay between an advertiser and the users is modeled under a contact, in which different user types are devised to account for the users' enjoyment, disutility, and ad sensitivity, respectively. In [19], the contractual agreement between a federated learning model owner and different Unmanned Aerial Vehicles (UAVs) that offer their computation capabilities is examined, in which each UAV is jointly distinguished based on its sensing, computation, and transmission costs. Last, in [20] the problem of optimal wireless data plans offered by a Mobile Network Operator (MNO) to its subscribing users is studied, by incorporating the users' satisfaction and network substitutability as two distinct user types in the model.

With reference to the computation offloading under single-layer computing environments, a wide variety of works exist in the literature. Indicative ones in [21], [22], treat unilaterally the problem of computation offloading from different perspectives, accounting for multi-server setups [21] or devising usage-based pricing policies [22]. Other attempts, e.g., [23]–[25], focus on the challenging joint communication and computing resource allocation under NOMA-enabled computing systems, by mainly proposing game-theoretic approaches to obtain a solution in tractable manner and within polynomial time [26]. In [23], the authors aim to minimize the users' sum delay by optimizing their offloading strategies and uplink transmission powers to the edge server. Optimizing a similar set of variables, the minimization of the total energy is pursued in [24], while the concurrent minimization of the users' energy consumption and latency is achieved in [25], via a Stackelberg game. However, all aforementioned works in [21]–[25], consider this single layer as a practically infinite energy and resource computing layer compared to the users' constrained devices, whereas our approach removes this limitation, by taking the edge service layer's energy efficiency into account.

B. CONTRIBUTIONS AND OUTLINE

It becomes apparent that although several efforts have been devoted to the joint communication and computing resource allocation that pertain to different multi-layer computing settings, the overwhelming majority of them is founded on the effective and efficient execution of delay-sensitive tasks. In our paper, in contrast to the rest of the research works, we aim to, first, study the problem of collaborative edge-fog computing from a market perspective and leverage on the economic interplay between the involved parties to tackle the challenging two-layer computing environment's resource orchestration. Under this objective, our goal is to better utilize the available computing resources in such a heterogeneous and multi-layer computing setting, increasing in this way its computing service capacity, while minimizing the end-to-end energy overhead. Specifically, the key contributions of this paper are summarized as follows.

1) A system model of a two-layer edge-fog computing environment is introduced, accounting for both the computing models of the users, the edge and fog servers, and the wireless users-to-edge and edge-to-fog communication models (Section II).
2) An incentive mechanism is designed between the edge server and the users following the principles of multi-dimensional contract theory. Based on the heterogeneity of the users' applications and hence, their multi-dimensional private information, the edge server derives a set of contract bundles, comprising the required efforts from the users and their offered rewards. Each user's effort represents the percentage of the initially offloaded task at the edge server that can be further transmitted and processed at the fog (Section III).
3) A joint communication and computing resource allocation problem is designed between the edge server and the users in the form of a Stackelberg game. The edge server, i.e., the leader, determines the users' optimal amounts of tasks to be offloaded at the edge, being aware of the percentage of each user's task that will be processed at the fog. Subsequently, the users, i.e., the followers, being multiplexed via power-domain NOMA, derive their optimal uplink transmission powers to the edge. The edge server seeks to maximize

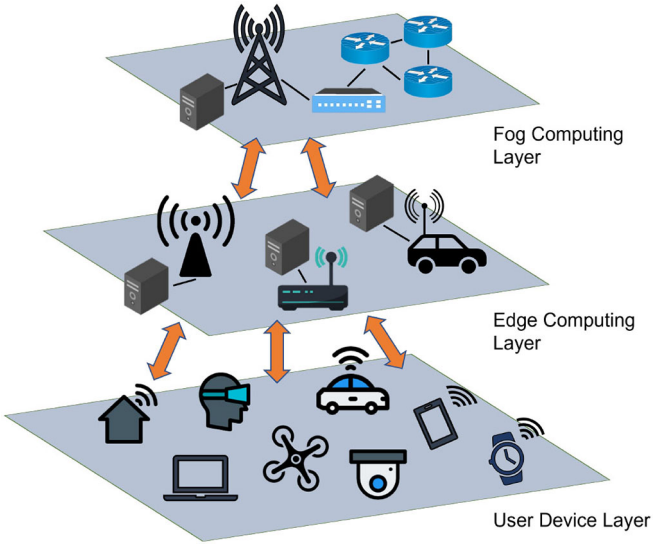

FIGURE 1. High-level overview of the two-layer computing environment's architecture.

its perceived satisfaction minus the end-to-end energy overhead from the users to the fog, while the users pursue their personal energy efficiency maximization under a non-cooperative game. The overall resource allocation procedure is iteratively executed until the Stackelberg equilibrium is reached (Section IV).

4) Based on the above theoretical foundations, we study the inherent operational characteristics of both the incentive mechanism and the resource allocation procedure, via modeling and simulation. Moreover, we prove the performance efficiency of the proposed incomplete information contract by comparison with the benchmark complete information case, while demonstrating, at the same time, the superiority of the proposed resource allocation approach, against different baseline offloading strategies (Section V).

II. SYSTEM MODEL

A two-layer computing environment is considered, consisting of a set of users $\mathcal{N} = \{1, \ldots, N\}$, an edge server, and a fog server. We assume that the edge server can be mobile, with consequently some limitation on its available energy, and hence, can move in close proximity to the users. On the other hand, the fog server lies between the edge and the cloud/core network, serving - among others - the purpose of computation alleviation/relaxation of the edge. It should be noted that the problem of the edge server placement, though interesting and challenging, is considered beyond the scope of the paper, while the extension to the multi-edge server case is part of our future work. The focus of this paper is primarily placed on the interplay between the various computing layers (users-edge-fog) and their joint and collaborative exploitation. A high-level overview of the general users-edge-fog computing architecture, aligned with the system model considered in this paper, is presented in Fig. 1. Fig. 1 highlights the architectural differentiation between the edge and fog computing layers, with respect to the location where their intelligence and computation power is placed within the overall network [27]–[30].

In this system, each user n has a computing application A_n, which can range from a typical smart city, transportation, healthcare, industry and agriculture computing application (e.g., [1], [31]), as illustrated in Fig. 1. Each user's computing application's A_n specific characteristics are defined as $A_n = (D_n, \phi_n, T_n, E_n)$, where D_n [Bytes] denotes the application's total input bytes, ϕ_n [CPU cycles/Byte] indicates the application's intensity and T_n [s] is the end-to-end completion time requirement, which implicitly reveals the user's level of delay tolerance. Last, E_n [J] is the user device's energy constraint. Accordingly, the term $\phi_n D_n$ [CPU cycles] denotes the number of CPU cycles required for the application's execution, which is referred to as "task" in the following and can represent a number of images, videos, text, voice, or maps, depending on the user's computing application's nature. In this paper, we pursue a realistic scenario, under which the user application's characteristics take values from discrete sets, such that $D_n \in \mathcal{D}$, $\phi_n \in \Phi$, $T_n \in \mathcal{T}$ and $E_n \in \mathcal{E}$, where $\mathcal{D}, \Phi, \mathcal{T}, \mathcal{E}$ are the corresponding discrete sets. Also, we assume that a task $\phi_n D_n$ can be arbitrarily partitioned into subsets of any size, which can be executed at either the user device, edge server, or fog server.

Owing to the edge server's appealing attributes, including its proximity to the users, we assume that each user n chooses to communicate with the edge server and offload part of its total task $\phi_n D_n$ for remote computation. We denote as $\phi_n d_n$ [CPU cycles] the part of task that is actually offloaded by the user at the edge, where $d_n \in [0, D_n]$ [Bytes] is the user's n offloading bytes. Based on the application's A_n characteristics, a percentage $x_n \in [0, 1]$ of the initially offloaded task $\phi_n d_n$ by the user n at the edge, is allowed to be further transmitted and processed at the fog. The value of the percentage x_n is derived from the contractual agreement between the edge server and the user n, which is analytically presented later in Section III. As a result, considering a user n, a total amount of $x_n d_n$ [Bytes] is wirelessly transmitted from the edge to the fog, and $x_n \phi_n d_n$ [CPU cycles] are computed at the fog server, while the remaining $(1 - x_n)\phi_n d_n$ are ultimately processed at the edge. Finally, it is noted that $D_n - d_n$ bytes are reserved for local computation at the user's device.

A. WIRELESS COMMUNICATION MODEL

Focusing on the communication model, we assume that the two-layer wireless network operates in out-of-band mode, meaning that the transmissions in the wireless access and backhaul network parts (e.g., user-to-edge and edge-to-fog) are performed using different frequency bands. We denote as W_e [Hz] the bandwidth of the wireless access and W_f [Hz] the bandwidth of the wireless backhaul that facilitates the transmission from the edge to the fog. The users' transmissions in the wireless access are multiplexed using the combination of power-domain NOMA

and Successful Interference Cancellation (SIC) techniques, while no interference is sensed by the edge server at the wireless backhaul network part.

In detail, regarding the wireless access of the users to the edge server, we denote by G_n the channel gain of a user n, which is defined as $G_n = \rho d_{n,e}^{-a_e}$, where ρ [dB] is the path loss at the reference distance of 1m, $d_{n,e}$ [m] is the Euclidean distance between the user n and the edge server and a_e is the path loss exponent. Without loss of generality, we assume that the users' channel gains are ordered in ascending manner, i.e., $G_1 \leq \cdots \leq G_n \cdots \leq G_N$, such that the decoding starts from the higher channel gain user, when the SIC technique takes place at the receiver of the edge server. Hence, following the combination of NOMA and SIC, the user's n achieved data rate in the uplink direction to the edge server is:

$$R_n = W_e \log_2\left(1 + \frac{G_n p_n}{\sum_{n'=1}^{n-1} G_{n'} p_{n'} + I_0}\right)[bps], \quad (1)$$

where $p_n \in [0, p_n^{max}]$ [W] indicates the user's n uplink transmission power that is constrained by a maximum transmission power level p_n^{max}, and I_0 [dBm/Hz] is the power spectral density of zero-mean Additive White Gaussian Noise (AWGN). As a result, considering that a user n transmits d_n bytes to the edge server, we can define the transmission time and energy overheads that experiences as follows:

1) User's n offloading time overhead:

$$T_n^{off} = \frac{d_n}{R_n}[s]. \quad (2)$$

2) User's n offloading energy overhead:

$$E_n^{off} = \frac{d_n p_n}{R_n}[J]. \quad (3)$$

Regarding the wireless backhaul transmission from the edge to the fog, we define as $G_e = \rho d_{e,f}^{-a_f}$ the channel gain between the edge and fog servers, where $d_{e,f}$ [m] is the Euclidean distance between the two servers and a_f is the corresponding path loss exponent. Denoting as p_e [W] the uplink transmission power of the edge server to the fog, the edge server's achieved data rate is expressed as:

$$R_e = W_f \log_2\left(1 + \frac{G_e p_e}{I_0}\right)[bps]. \quad (4)$$

Accordingly, we define the transmission time and energy overheads at the backhaul, experienced by the edge server, considering a single user n:

1) Edge server's offloading time overhead for user's n task:

$$T_e^{n,off} = \frac{x_n d_n}{R_e}[s]. \quad (5)$$

2) Edge server's offloading energy overhead for user's n task:

$$E_e^{n,off} = \frac{x_n d_n p_e}{R_e}[J]. \quad (6)$$

B. COMPUTING MODEL

The two-layer edge-fog computing setting under consideration provides, apparently, three levels of different computing capabilities to its serving users. Analyzing these options from a bottom-up perspective, we denote as F_n [CPU cycles/s] the user n device's inherent (local) computing capability, and σ_n [J/CPU cycle] its corresponding power consumption coefficient per CPU cycle. Considering that the user n executes locally a task of size $\phi_n(D_n - d_n)$ CPU cycles, the user's n corresponding computation/execution time and energy overheads are defined as follows [32], [33]:

1) User's n execution time overhead:

$$T_n^{exec} = \frac{\phi_n(D_n - d_n)}{F_n}[s]. \quad (7)$$

2) User's n execution energy overhead:

$$E_n^{exec} = \sigma_n \phi_n (D_n - d_n) F_n^2 [J]. \quad (8)$$

Adopting a similar modeling for the subsequent computing layer, we indicate as F_e [CPU cycles/s] the edge server's computing capability, which is assumed to be higher than each user's n, but finite and more limited compared to the fog. Also, let σ_e [J/CPU cycle] denote its power consumption coefficient per CPU cycle. Considering that the edge servers' resources are sufficient and can facilitate the parallel computation of the users' tasks [34], [35], the edge server's incurred time and energy consumption overheads for each user n are:

1) Edge server's execution time overhead for user's n task:

$$T_e^{n,exec} = \frac{(1 - x_n)\phi_n d_n}{F_e}[s]. \quad (9)$$

2) Edge server's execution energy overhead for user's n task:

$$E_e^{n,exec} = \sigma_e (1 - x_n)\phi_n d_n F_e^2 [J]. \quad (10)$$

As far as the fog computing layer is concerned, in this paper, we assume that the processing capabilities of the fog server significantly excel both the edge server and the user devices, and hence, without loss of generality and for simplicity in the presentation, we assume that the fog induces practically zero time and energy costs compared to the other two lower computing layers. However, this analysis is still valid and can be easily extended to additionally account for these overheads at the fog computing layer.

C. OVERALL FRAMEWORK

In this paper, we aim to promote the utilization of the end-to-end users-edge-fog computing paradigm, under the case that the users' tasks are characterized by some form of delay tolerance. Considering that the users typically exhibit a selfish and greedy behavior in relation to their perceived satisfaction from the remote computation of their tasks, we first employ an incentive mechanism that targets to shift their preference from the prevailing edge server to the fog server. To this end,

a multi-dimensional contract is designed by the edge server, which based on the edge server's statistical knowledge of the potential user applications' heterogeneous features, concludes to a set of optimal contract bundles $\mathbf{w}^* = \{\mathbf{x}^*, \mathbf{r}^*\}$ intended for the users. The term $\mathbf{x}^*$ indicates the vector of the users' required efforts, which are mapped to the percentage of each user's task $\phi_n d_n$ that is further offloaded at the fog, while the term $\mathbf{r}^*$ represents the vector of the corresponding rewards, which can be considered as a form of discount to the users' received computing service from the edge server. The exact definition of these parameters is provided later in Section III-A. Based on its private information, i.e., user application characteristics, each user autonomously selects the contract bundle $w_n^* = \{x_n^*, r_n^*\}$ that best fits is type, revealing implicitly in this way its private information to the edge server.

Being aware of the percentage of task x_n^* that is allowed by each user n to be transmitted to the fog, the edge server is able to calculate the end-to-end system's total energy and time overhead from the users to the fog, as presented in Eq. (2)-(3) and Eq. (5)-(10). Targeting to maximize its perceived satisfaction minus the end-to-end computing environment's energy overhead, the edge server calculates each user's n optimal amount of offloaded task $\phi_n d_n^*$, having a prior knowledge of the users' uplink transmission powers to the edge. The edge server's decision if fed back to the users, who accordingly determine their optimal uplink transmission power levels $p_n^*, \forall n \in \mathcal{N}$ in a distributed manner, by participating in a non-cooperative game among themselves. The optimization procedures of both the edge server and the users are iteratively updated, until convergence of this joint resource allocation is achieved, resulting in a Stackelberg game. A high-level overview of the individual steps of the proposed incentive mechanism and resource allocation framework is presented in Fig. 2, revealing also the interactions that take place between the edge server and the users at each step.

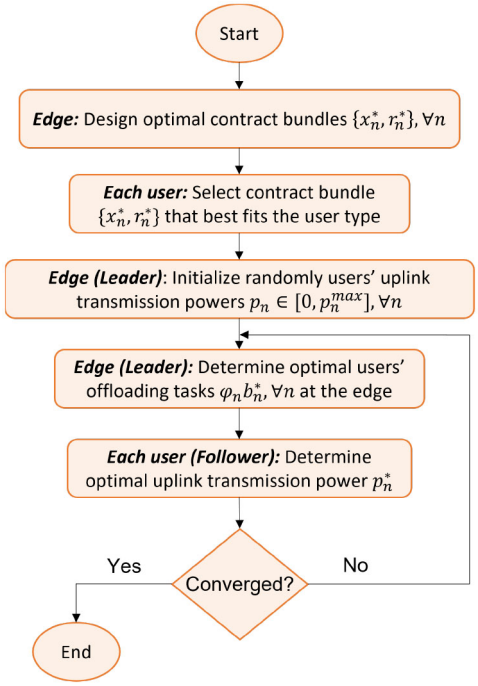

FIGURE 2. High-level overview of the overall incentive mechanism and resource allocation framework.

III. MULTI-DIMENSIONAL CONTRACT-BASED INCENTIVE MECHANISM DESIGN

This section is devoted to the introduction and analytic description of the devised incentive mechanism, based on multi-dimensional contract theory. First, we define the multi-dimensional private information and user types that distinguish the users and reflect their heterogeneity, along with the designed contract bundles and utilities. Then, we study the problem of contract formulation and gradually, we derive the optimal contracts.

A. USER TYPES, CONTRACT BUNDLES & UTILITIES

Following each user application's A_n heterogeneous characteristics, the users can be categorized into different user types that capture their ability and willingness to allow part of their initially offloaded tasks at the edge server to be forwarded to the fog. Assuming that the discrete sets $\mathcal{D}, \Phi, \mathcal{T}$ comprise K, L, M values, respectively, such that $\mathcal{D} = \{\mathcal{D}_k : 1 \leq k \leq K\}$, $\Phi = \{\Phi_l : 1 \leq l \leq L\}$ and $\mathcal{T} = \{\mathcal{T}_m : 1 \leq m \leq M\}$, there exist $K \times L \times M$ combinations of user types in the system, which derive from the Cartesian product $B \times \Gamma \times \Delta$ of the sets B, Γ, Δ analyzed in the following. Specifically, we categorize the users into a set $B = \{\beta_k : 1 \leq k \leq K\}$ of K application size evaluation types, which are determined by the rule $\beta_k = \frac{\mathcal{D}_k}{\max_{1 \leq k \leq K}\{\mathcal{D}_k\}}$, a set $\Gamma = \{\gamma_l : 1 \leq l \leq L\}$ of L application intensity evaluation types that are derived as $\gamma_l = \frac{\Phi_l}{\max_{1 \leq l \leq L}\{\Phi_l\}}$, as well as a set $\Delta = \{\delta_m : 1 \leq m \leq M\}$ of M delay sensitivity evaluation types, such that $\delta_m = \frac{1/\mathcal{T}_m}{\max_{1 \leq m \leq M}\{1/\mathcal{T}_m\}}$. For all user types it holds that $\beta_k, \gamma_l, \delta_m \in (0, 1], \forall k, l, m$. Without loss of generality, we assume that the user types are sorted in ascending order under all dimensions, i.e., $\beta_1 \leq \cdots \leq \beta_k \leq \cdots \leq \beta_K$, $\gamma_1 \leq \cdots \leq \gamma_l \leq \cdots \leq \gamma_L$, $\delta_1 \leq \cdots \leq \delta_m \leq \cdots \leq \delta_M$. Also, each combination of the $K \times L \times M$ user types is characterized by a joint probability mass function $Pr(\beta_k, \gamma_l, \delta_m), \forall k, l, m$. At this point, it should be reminded that the user types constitute the users' private information that is unknown to the edge server, whereas the edge server is only aware of the different user types' joint probability mass function, and should appropriately design the contract bundles relying on this partially complete (or incomplete) information.

As mentioned earlier, the edge server designs a set of optimal contract bundles $\mathbf{w}^* = \{\mathbf{x}^*, \mathbf{r}^*\}$ of the users' required efforts and offered rewards, respectively, based on its probabilistic knowledge of the potential users' types. Specifically, we denote as $x_{k,l,m}^n \in [0, 1]$ the effort of user n of type $(\beta_k, \gamma_l, \delta_m)$, and $r_{k,l,m}^n \in \mathbb{R}^+$ its corresponding reward. Apparently, different users from the set $\mathcal{N}$ can be of the

same type $(\beta_k, \gamma_l, \delta_m)$, acquire the same contract bundle and hence, experience the same utility. Since our analysis is focused on the differentiation of the contract bundles with respect to the different user types, in the following, we drop the superscript n that points to a specific user for notation simplicity. In addition, we refer to the user type $(\beta_k, \gamma_l, \delta_m)$ as (k, l, m)-type user, whose corresponding contract bundle is $w_{k,l,m} = \{x_{k,l,m}, r_{k,l,m}\}$.

Given a (k, l, m)-type user's effort $x_{k,l,m}$ and its offered reward $r_{k,l,m}$, we define its perceived satisfaction from its participation in the contract by the following utility:

$$U_{k,l,m}\big(w_{k,l,m}\big) = r_{k,l,m} - (1.5 + \delta_m - \beta_k\gamma_l)q\big(x_{k,l,m}\big), \quad (11)$$

where $q(x_{k,l,m})$ is an increasing function of $x_{k,l,m}$, which together with the term $(1.5 + \delta_m - \beta_k\gamma_l)$ implies its evaluation of provided effort. Specifically, the term $\beta_k\gamma_l q(x_{k,l,m})$ indicates that the user's benefit from exerting its effort to the edge server increases as its β_k, γ_l types increase, since a higher amount of task can be executed at the combination of edge-fog. On the contrary, the user's benefit decreases as its delay sensitivity evaluation δ_m increases, noted by the term $-\delta_m q(x_{k,l,m})$. The physical meaning and interpretation of the overall utility function is that the (k, l, m)-type user's satisfaction derives from its offered reward minus its provided effort, which, in turn, increases proportionally to its application's input bytes and intensity and is inversely proportional to its delay sensitivity. For ease of reference, we define as $Q_{k,l,m}(x_{k,l,m}) = -(1.5 + \delta_m - \beta_k\gamma_l)q(x_{k,l,m})$, and we rewrite the (k, l, m)-user type's utility as $U_{k,l,m}(w_{k,l,m}) = Q_{k,l,m}(x_{k,l,m}) + r_{k,l,m}$. Taking into account that $\beta_k, \gamma_l, \delta_m \in (0, 1]$ and that $q(x_{k,l,m})$ is an increasing function, it holds that $\frac{\partial Q_{k,l,m}}{\partial x_{k,l,m}} < 0$, and thus, $Q_{k,l,m}(x_{k,l,m})$ is a decreasing function on $x_{k,l,m}$. For demonstration purposes and without loss of generality, in the following we consider $q(x_{k,l,m}) = x_{k,l,m}^2$.

Concerning the utility that the edge server attains from a single (k, l, m)-type user's participation in the contract, this is modeled as $V_{k,l,m}(w_{k,l,m}) = h(x_{k,l,m}) - \xi r_{k,l,m}$, where $h(x_{k,l,m})$ is an increasing and concave function on $x_{k,l,m}$, accounting for the edge server's evaluation of received effort, while $\xi \geq 1$ is the edge server's cost of offered rewards. Obviously, the edge server's utility increases as the user's effort increases and decreases proportionally to its offered reward. As a result, the overall edge server's utility from the different user types' participation can be expressed as follows:

$$V(W) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} Pr_{k,l,m}\big(h(x_{k,l,m}) - \xi r_{k,l,m}\big), \quad (12)$$

where $W = \{w_{k,l,m}, 1 \leq k \leq K, 1 \leq l \leq L, 1 \leq m \leq M\}$ is the set that contains the ensemble of contract bundles. Motivated by [18], in the following we consider the function $h(x_{k,l,m}) = \frac{c}{1-\lambda}x_{k,l,m}^{1-\lambda}$ to capture the sharp increase of the edge server's marginal rate of satisfaction with the user's effort increase, by intelligently controlling $c \in [0, 1]$ and $\lambda \in (0, 1)$.

B. CONTRACT FORMULATION

Considering the realistic scenario of incompleteness of information from the edge server's perspective, then the designed contract bundles should bear specific properties in order to promote the users' participation in the contract. In particular, the edge server should ensure that each user experiences a non-negative utility, while its utility is, also, maximized when selecting the contract bundle designated for its specific type. These two conditions are summarized under the notions of Individual Rationality (IR) and Incentive Compatibility (IC), which are formally defined below.

Definition 1 [Individual Rationality (IR)]: A contract bundle $w_{k,l,m} = \{x_{k,l,m}, r_{k,l,m}\}$ satisfies the individual rationality condition if each (k, l, m)-type user for all $1 \leq k \leq K, 1 \leq l \leq L, 1 \leq m \leq M$ receives a non-negative utility, i.e.,

$$U_{k,l,m}\big(w_{k,l,m}\big) \geq 0, \forall k, l, m. \quad (13)$$

Definition 2 [Incentive Compatibility (IC)]: Each (k, l, m)-type user for all $1 \leq k \leq K, 1 \leq l \leq L, 1 \leq m \leq M$ receives the maximum utility, when selecting the contract bundle $w_{k,l,m} = \{x_{k,l,m}, r_{k,l,m}$ that is intended for its own type, i.e.,

$$U_{k,l,m}\big(w_{k,l,m}\big) \geq U_{k,l,m}\big(w_{k',l',m'}\big), \forall k, l, m, \quad k \neq k', l \neq l', m \neq m'. \quad (14)$$

Hence, the multi-dimensional contract problem to be solved by the edge server can be formally written as:

$$\max_{W} \; V(W) \quad \textbf{s.t.} \; (13), (14). \quad (15)$$

The resulting optimization problem in Eq. (15) includes KLM IR and $KLM(KLM - 1)$ IC constraints, which fully interconnect the contract bundle design between the different user types. In order to derive a tractable solution, an appropriate procedure should take place to reduce its constraints, which primarily differs from the standard method used in the one-dimensional contract problems (e.g., [10]–[12], [14]–[17]) and is comprehensively presented in Sections III-C and III-D.

C. CONTRACT FEASIBILITY

In this section, we study the necessary conditions that must be satisfied in order to render the formulated contract problem feasible. To facilitate this analysis, we first transform the three-dimensional contract problem to a single-dimensional one, by introducing a single "virtual" user type that bears all three-dimensional private information of the users. To this end, we resort to some useful properties of the theory of economics.

We consider a (k, l, m)-type user's indifference curve in the contract plane $\{x_{k,l,m}, r_{k,l,m}\}$ between its effort and reward, which under a fixed utility value $U(w) = \bar{U}$ satisfies:

$$\bar{U} = r_{k,l,m} - (1.5 + \delta_m - \beta_k\gamma_l)q\big(x_{k,l,m}\big) = Q_{k,l,m}\big(x_{k,l,m}\big) + r_{k,l,m}, \quad (16)$$

which actually yields all combinations of $\{x_{k,l,m}, r_{k,l,m}\}$ that result in the same utility to the users.

The slope s of the indifference curve is calculated by taking the partial derivatives of both sides in Eq. (16) as:

$$s_{k,l,m}\big(x_{k,l,m}\big) = -\frac{\partial Q_{k,l,m}}{\partial x_{k,l,m}} = \frac{\partial r_{k,l,m}}{\partial x_{k,l,m}} = (1.5 + \delta_m - \beta_k\gamma_l)q'\big(x_{k,l,m}\big), \quad (17)$$

which is referred to as "marginal rate of substitution", implying the rate at which a user is expected to abandon a $\{x_{k,l,m}, r_{k,l,m}\}$ combination in exchange for another, while maintaining the same utility value. Apparently, Eq. (17) depends on the three-dimensional (k, l, m)-type in a combined manner, and the effort $x_{k,l,m}$. By scrutinizing the definition of parameter s, we can easily deduce that the lower the value of s, then the lower the delay sensitivity evaluation type δ and the higher the application size and intensity evaluation types β and γ, respectively. This concludes in higher user's willingness to participate in the contract. On the contrary, the opposite holds true under a higher value of s, which increases the user's unwillingness to participate. Hence, as the value of the marginal rate of substitution s in Eq. (17) increases, the user's overall willingness decreases. Therefore, we may additionally refer to s as "unwillingness-to-participate" parameter.

Without loss of generality, we sort the $K \times L \times M$ user types in ascending order with respect to the unwillingness-to-participate parameter s as follows:

$$Z_1(x), \ldots, Z_i(x), \ldots, Z_{KLM}(x), \quad (18)$$

where $Z_i(x) \triangleq (\beta_i, \gamma_i, \delta_i)$, $1 \leq i \leq KLM$ denotes a user type under the new formulation, which is actually the virtual user type we are seeking. Thus, considering an effort x and under the ordering in Eq. (18), it holds that:

$$s(Z_1, x) \leq \cdots \leq s(Z_i, x) \leq \cdots \leq s(Z_{KLM}, x). \quad (19)$$

It is interesting that although the value of $s(Z_i, x)$ changes for different efforts x, the virtual user type ordering in Eq. (18) remains unchanged, which we elaborate on Lemma 1 below.

Lemma 1: The new user type ordering in Eq. (19) is independent of the effort x, i.e., $Z_i(x) = Z_i(x')$, $x \neq x'$, $1 \leq i \leq KLM$.

Proof: The proof of this lemma stems intuitively from the fact that the unwillingness-to-participate parameter $s(\beta, \gamma, \delta, x)$ has a separable structure with respect to the three-dimensional types (β, γ, δ) and the effort x, as can be easily observed by its definition in Eq. (17). ■

In the remainder of the paper, we directly use Z_i to refer to the virtual user type, referred to as unwillingness-to-participate user type. Also, the contract bundle intended for Z_i is denoted as $w_i = \{x_i, r_i\}$, while this notation applies to all other considered variables. Consequently, the utility function of a user type Z_i is written as $U_i = Q(Z_i, x_i) + r_i$, where $Q(Z_i, x_i)$ is the equivalent of $Q_{k,l,m}(x_{k,l,m})$.

Given the outcome of Lemma 1, we conclude that whatever the value of x is, the minimum unwillingness-to-participate user type is $Z_1 = (\beta_K, \gamma_L, \delta_1)$, whose application size and intensity is the highest, while its delay sensitivity is the lowest. Conversely, the maximum unwillingness-to-participate user type is $Z_{KLM} = (\beta_1, \gamma_1, \delta_M)$, which, also, attains the minimum utility based on the definition of utility in Eq. (11).

Next, we derive the necessary conditions that render a contract $W = \{w_i, 1 \leq i \leq KLM\}$ feasible, meaning that the IR and IC conditions defined in Eq. (13) and Eq. (14), respectively, are successfully met.

Lemma 2: For any feasible contract $W = \{w_i, 1 \leq i \leq KLM\}$, it holds true that $x_i > x_j \Leftrightarrow r_i > r_j$.

Proof: First, we prove that $r_i > r_j \Rightarrow x_i > x_j$, by utilizing the IC condition that holds true for user type Z_j, which gives $Q(Z_j, x_j) + r_j \geq Q(Z_j, x_i) + r_i \Leftrightarrow Q(Z_j, x_j) - Q(Z_j, x_i) \geq r_i - r_j$. Thus, if $r_i > r_j$ then $Q(Z_j, x_j) > Q(Z_j, x_i)$ and considering that function Q is decreasing with respect to x, we get $x_i > x_j$.

In order to prove that $x_i > x_j \Rightarrow r_i > r_j$, we follow a similar procedure and we elaborate on the IC condition that holds for user type Z_i as: $Q(Z_i, x_i) + r_i \geq Q(Z_i, x_j) + r_j \Leftrightarrow Q(Z_i, x_i) - Q(Z_i, x_j) \geq r_j - r_i$. Then, if $x_i > x_j \overset{Q\searrow}{\Longleftrightarrow} Q(Z_i, x_i) < Q(Z_i, x_j)$ and thus, it can be easily concluded that $r_i > r_j$. This completes the proof. ■

The rationale behind Lemma 2 is that a user receives a higher reward, when providing a higher effort to the edge server, in order to be properly incentivized to participate in the contract.

Lemma 3 (Monotonicity): For any feasible contract $W = \{w_i, 1 \leq i \leq KLM\}$, it holds true that $s(Z_i, x) > s(Z_j, x) \Rightarrow x_i \leq x_j$, for any x.

Proof: We prove this lemma by contradiction, assuming that there exist x_i and x_j, such that $x_i > x_j$, which give $s(Z_i, x) > s(Z_j, x)$, for any x.

We write the IC conditions that hold for the user types Z_i and Z_j, respectively, as $Q(Z_i, x_i) + r_i \geq Q(Z_i, x_j) + r_j$ and $Q(Z_j, x_j) + r_j \geq Q(Z_j, x_i) + r_i$. By adding these two IC condition inequalities by parts, we get $Q(Z_i, x_i) + Q(Z_j, x_j) \geq Q(Z_i, x_j) + Q(Z_j, x_i)$, which is equivalently written as:

$$\big[Q(Z_i, x_i) + Q\big(Z_j, x_j\big)\big] - \big[Q\big(Z_i, x_j\big) + Q\big(Z_j, x_i\big)\big] \geq 0. \quad (20)$$

Elaborating on Eq. (20) via the fundamental theorem of calculus, we obtain:

$$\begin{aligned}&\big[Q(Z_i, x_i) + Q\big(Z_j, x_j\big)\big] - \big[Q\big(Z_i, x_j\big) + Q\big(Z_j, x_i\big)\big] \\ &= \int_{x_j}^{x_i} \frac{\partial Q(Z_i, x)}{\partial x} dx - \int_{x_j}^{x_i} \frac{\partial Q\big(Z_j, x\big)}{\partial x} dx \\ &= \int_{x_j}^{x_i} -\big[s(Z_i, x) - s\big(Z_j, x\big)\big] dx. \end{aligned} \quad (21)$$

Since $x_i > x_j$, Eq. (21) gives $s(Z_i, x) < s(Z_j, x)$, which contradicts with our initial assumption. In this way, we have proved that there does not exist $x_i > x_j$ such that

$s(Z_i, x) > s(Z_j, x)$, which confirms the soundness of this lemma. ■

The reasoning behind the monotonicity condition in Lemma 3 is that a higher unwillingness-to-participate user type provides a lower effort to the edge server and thus, is rewarded less, taking also into account Lemma 2.

Based on the above analysis, we can summarize the necessary conditions of a feasible contract in the following theorem.

Theorem 1 (Necessary Conditions): A feasible contract $W = \{w_i, 1 \leq i \leq KLM\}$ must meet the following two conditions concurrently: $x_1 \geq \cdots \geq x_i \geq \cdots \geq x_{KLM}$ and $r_1 \geq \cdots \geq r_i \geq \cdots \geq r_{KLM}$.

D. CONTRACT SUFFICIENCY

In this section, we resolve the problem of reducing the IR and IC conditions, defined in Eq. (13) and Eq. (14), respectively, as a means of obtaining and optimal solution for the contract problem designed in Eq. (15) in Section III-B. The outcome of this analysis is the definition of the sufficient conditions of a feasible contract under the realistic scenario of incompleteness of information.

Lemma 4 (IR Conditions Reduction): Under a feasible contract, if the IR condition of the lowest utility user type, i.e., the highest unwillingness-to-participate user type Z_{KLM}, holds true, then the IR conditions of all other user types are automatically satisfied:

$$U_{KLM}(w_{KLM}) \geq 0 \Leftrightarrow U_i(w_i) \geq 0, 1 \leq i \leq KLM. \quad (22)$$

Proof: The IC condition that holds between a user type $Z_i, 1 \leq i \leq KLM$ and the lowest utility user type Z_{KLM} is $U_i(w_i) \geq U_i(w_{KLM})$. Furthermore, for the minimum utility user type it holds that $U_{KLM}(w_{KLM}) \leq U_i(w_{KLM}), 1 \leq i \leq KLM$. Thus, if $U_{KLM}(w_{KLM}) \geq 0$, then $U_i(w_i) \geq 0, 1 \leq i \leq KLM$. This completes the proof. ■

The Lemma 4 allows the reduction of the KLM IR constraints of the optimization problem in Eq. (15) to a single IR constraint, i.e., $U_{KLM}(w_{KLM}) \geq 0$.

Next, we introduce the Pairwise Incentive Compatibility (PIC) condition to facilitate the IC conditions reduction process later in this section.

Lemma 5: Pairwise Incentive Compatibility (PIC): The contract bundles $w_i, w_j \in W, 1 \leq i, j \leq KLM, i \neq j$ are pairwise incentive compatible, denoted as $w_i \overset{PIC}{\Longleftrightarrow} w_j$, if the following two conditions are concurrently satisfied: $U_i(w_i) \geq U_i(w_j)$ and $U_j(w_j) \geq U_j(w_i)$.

Lemma 6: IC Conditions Reduction: Under a feasible contract, the following condition holds true for any $i_1 < i_2 < i_3$:

$$\text{If } w_1 \overset{PIC}{\Longleftrightarrow} w_2 \text{ and } w_2 \overset{PIC}{\Longleftrightarrow} w_3, \text{ then } w_1 \overset{PIC}{\Longleftrightarrow} w_3. \quad (23)$$

Proof: To prove this lemma, we write the IC conditions that are satisfied for the user types Z_{i1} and Z_{i2}, respectively, as:

$$Q(Z_{i_1}, x_{i_1}) + r_{i_1} \geq Q(Z_{i_1}, x_{i_2}) + r_{i_2}, \quad (24)$$

and

$$Q(Z_{i_2}, x_{i_2}) + r_{i_2} \geq Q(Z_{i_2}, x_{i_3}) + r_{i_3}. \quad (25)$$

Since $i_1 < i_2 < i_3$, from Theorem 1 we have $x_{i_1} > x_{i_2} > x_{i_3}$ and $s(Z_{i_1}, x) < s(Z_{i_2}, x) < s(Z_{i_3}, x)$. Founded upon this, it holds that $\int_{x_{i_3}}^{x_{i_2}} [s(Z_{i_2}, x) - s(Z_{i_1}, x)]dx \geq 0$, on which we subsequently elaborate according to the fundamental theorem of calculus as:

$$\begin{aligned}
&\int_{x_{i_3}}^{x_{i_2}} \big[s(Z_{i_2}, x) - s(Z_{i_1}, x)\big]dx \\
&= \int_{x_{i_2}}^{x_{i_3}} \frac{\partial Q(Z_{i_2}, x)}{\partial x}dx - \int_{x_{i_2}}^{x_{i_3}} \frac{\partial Q(Z_{i_1}, x)}{\partial x}dx \\
&= \big[Q(Z_{i_2}, x_{i_3}) - Q(Z_{i_2}, x_{i_2})\big] - \big[Q(Z_{i_1}, x_{i_3}) - Q(Z_{i_1}, x_{i_2})\big].
\end{aligned} \quad (26)$$

Therefore, the following condition holds, also, true:

$$Q(Z_{i_2}, x_{i_3}) - Q(Z_{i_2}, x_{i_2}) \geq Q(Z_{i_1}, x_{i_3}) - Q(Z_{i_1}, x_{i_2}). \quad (27)$$

By adding the three inequalities in Eq. (24), Eq. (25) and Eq. (27) by parts, we get:

$$Q(Z_{i_1}, x_{i_1}) + r_{i_1} \geq Q(Z_{i_3}, x_{i_3}) + r_{i_3}. \quad (28)$$

Another set of IC conditions that can be written for the user types Z_{i_3} and Z_{i_2}, is:

$$Q(Z_{i_3}, x_{i_3}) + r_{i_3} \geq Q(Z_{i_3}, x_{i_2}) + r_{i_2}, \quad (29)$$

and

$$Q(Z_{i_2}, x_{i_2}) + r_{i_2} \geq Q(Z_{i_2}, x_{i_1}) + r_{i_1}. \quad (30)$$

By applying similar steps to Eq. (26) via the use of the fundamental theorem of calculus, we can easily conclude to the next condition:

$$Q(Z_{i_3}, x_{i_2}) - Q(Z_{i_3}, x_{i_1}) \geq Q(Z_{i_2}, x_{i_2}) - Q(Z_{i_2}, x_{i_1}). \quad (31)$$

We add Eq. (29)-(31) by parts and we obtain:

$$Q(Z_{i_3}, x_{i_3}) + r_{i_3} \geq Q(Z_{i_3}, x_{i_1}) + r_{i_1}. \quad (32)$$

The combination of Eq. (28) and Eq. (32) proves that $w_1 \overset{PIC}{\Longleftrightarrow} w_3$, confirming the lemma. ■

The Lemma 6 enables the reduction of the $KLM(KLM-1)$ IC constraints of the optimization problem in Eq. (15) into a set of $2(KLM-1)$ PIC constraints between the adjacent user types Z_i and $Z_{i+1}, 1 \leq i \leq KLM-1$.

By combining our findings in Sections III-C and III-D so far, we can summarize the sufficient conditions for a feasible contract in the Theorem 2, below.

Theorem 2 (Sufficient Conditions): Under a feasible contract, the IR and IC conditions can be reduced as:

1) $x_1 \geq \cdots \geq x_i \geq \cdots \geq x_{KLM}$,
2) $U_{KLM}(w_{KLM}) \geq 0$,
3) $r_{i+1} - Q(Z_{i+1}, x_i) + Q(Z_{i+1}, x_{i+1}) \leq r_i \leq r_{i+1} + Q(Z_i, x_{i+1}) - Q(Z_i, x_i), 1 \leq i \leq KLM-1$.

The first condition in Theorem 2 derives from the necessary conditions in Theorem 1, while the second condition

stems from the findings in Lemma 4 and pertains to the IR conditions reduction. Concerning the third condition in Theorem 2, this represents the PIC conditions, as defined in Lemma 5, between two adjacent user types i and $i+1$, which constitutes the sufficient condition that should be met as a result of the IC conditions reduction procedure described in Lemma 6.

Based on the preceding analysis and the reduced IR and IC conditions listed in Theorem 2, the optimization problem in Eq. (15) is equivalently transformed as follows:

$$\max_{W} \; V(W) = \sum_{i=1}^{KLM} Pr_i(h(x_i) - \xi r_i)$$
$$\textbf{s.t.} \;\; \text{Conditions in Theorem 2.} \tag{33}$$

Without loss of generality and for ease in the optimal contract bundles' derivation, we consider the users' rewards as strictly increasing functions with their efforts in respect to the fact that the edge server acts fairly and rewards more the users that provide a higher effort, as described in Theorem 1, and we define $r_i(x_i) = \sqrt{Z_i}x_i$. Hence, the edge server's utility function $V(W)$, as described in Eq. (12), is concave as the sum of concave functions on x_i, $1 \leq i \leq KLM$, while the reduced constraints form a convex set. Thus, by applying standard optimization methods and utilizing existing concave/convex optimization tools [36], the optimal contract bundles $w_i^* = \{x_i^*, r_i^*\}$, $1 \leq i \leq KLM$.

E. BENCHMARK CONTRACT UNDER COMPLETE INFORMATION

In this section, a benchmark contract-based incentive mechanism is introduced, which is related to the case of complete information, in the sense that the edge server is a priori aware of the users' private information, i.e., their unwillingness-to-participate user types. Under this ideal case, the edge server can fully exploit the users' efforts and marginally satisfy their IR conditions to ensure their participation in the contract. Hence, the optimization problem to be solved by the edge server for each user type $1 \leq i \leq KLM$, is written as follows:

$$\max_{w_i} \; v_i = h(x_i) - \xi r_i, \; 1 \leq i \leq KLM \tag{34a}$$
$$\textbf{s.t.} \;\; Q(Z_i, x_i) + r_i = 0. \tag{34b}$$

From Eq. (34b) we get $r_i^* = -Q(Z_i, x_i) = Z_i x_i^2$, while Eq. (34a) is written as $v_i = h(x_i) - \xi r_i = \frac{c}{1-\lambda}x^{1-\lambda} - \xi r_i$, based on the provided definitions of the functions $Q(\cdot)$, $q(\cdot)$, $h(\cdot)$ earlier in this section. By substituting r_i^* back to v_i and calculating the first order derivative of v_i with respect to x_i, we get $\frac{\partial v_i}{\partial x_i} = \frac{c}{x_i^{\lambda}} - 2\xi Z_i x_i$. By solving the equation $\frac{\partial v_i}{\partial x_i} = 0$ with respect to x_i we obtain the optimal solution of the optimization problem, which is expressed as $x_i^* = (\frac{c}{2\xi Z_i})^{\frac{1}{1+\lambda}}$.

With reference to the feasibility of the optimization problem in Eq. (34a)-(34b), by taking into account that $\xi \geq 1$, $c \in [0, 1]$, $\lambda \in (0, 1)$ and $Z_i \in [0.5, 2.5]$, $1 \leq i \leq KLM$, the latter of which is determined by calculating the extreme values of the term $Z_i = 1.5 + \delta_i - \beta_i\gamma_i$, it holds that $x_i^* \geq 0$, $1 \leq i \leq KLM$. Additionally, considering the extreme case that $c = \xi = 1$ and $Z_i = 0.5$, under which the term $\frac{c}{2\xi Z_i}$ takes its highest value that is equal to $\frac{c}{2\xi Z_i} = 1$, it is verified that $x_i^* \leq 1$, $1 \leq i \leq KLM$. Hence, the optimal solution x_i^* is within the required range $[0, 1]$, yielding a feasible solution to the problem.

IV. STACKELBERG GAME-BASED RESOURCE ALLOCATION

After the completion of the multi-dimensional contract-based incentive mechanism, each user n has autonomously - and via the interaction with the edge server - determined its optimal amount of effort, i.e., the percentage x_n^* of the task $\phi_n d_n$ that is offloaded at the edge, which is allowed to be further transmitted and processed at the fog. Depending on the user's n application's A_n characteristics, each user n is represented by an unwillingness-to-participate user type Z_i and thus, the optimal contract bundle for this user is $w_n^* = \{x_n^*, r_n^*\} \leftrightarrow w_i^* \in W = \{w_i^*, 1 \leq i \leq KLM\}$.

At this second stage, the joint communications and computing resource allocation is pursued under a Stackelberg game-theoretic approach, in which the edge server (i.e., the leader) determines each user's n optimal offloaded task $\phi_n d_n^*$ and the users (i.e., the followers) decide on their optimal uplink transmission power p_n^* in an iterative manner, by exchanging information from one another. Specifically, with the term "task offloading optimization" we refer to the optimal amount of bytes d_n^* offloaded by each user n that determine the whole optimal amount of offloaded task $\phi_n d_n^*$. It should be noted that the joint computation task offloading and uplink transmission power allocation problem in NOMA-enabled computing environments, under both the energy efficiency maximization and the delay/time minimization objectives, as adopted in this paper and presented later in this section, is generally non-convex and NP-hard [23], [37], [38]. As a result, there does not exist any algorithm that provides an optimal solution to this joint problem in polynomial time. For this reason, either approximation or alternating [39] optimization algorithms are proposed in the literature to deal with it. Indeed, the Stackelberg game-theoretic approach proposed in this paper is aligned with both the decentralized and iterative optimization needs of the considered two-variable problem.

Next, the optimization problems of the leader and followers are presented, while the overall incentive mechanism and resource allocation framework is summarized in Algorithm 1.

A. LEADER'S OPTIMIZATION

Given the users' uplink transmission power vector $\mathbf{p} = [p_1, \ldots, p_n, \ldots, p_N]$, the edge server seeks to maximize its perceived satisfaction minus the end-to-end edge-fog computing environment's incurred energy overhead. To this end, the edge server determines the vector of the optimal amount of offloaded bytes $\mathbf{d}^* = [d_1^*, \ldots, d_n^*, \ldots, d_N^*]$, while meeting the users' end-to-end completion time and energy constraints,

and its personal energy constraint that stems from its inherent limitation on its available energy. Therefore, the corresponding optimization problem that is treated by the edge server is formulated as follows:

$$\max_{\mathbf{d}} \sum_{n=1}^{|N|}\Big[1 - e^{\frac{2d_n}{D_n}} - \mathcal{C}\Big(E_n^{off} + E_n^{exec} + E_e^{n,off} + E_e^{n,exec}\Big)\Big] \tag{35a}$$

$$\text{s.t. } 0 \leq d_n \leq D_n, \forall n \in \mathcal{N} \tag{35b}$$

$$\max\Big\{T_n^{off}, T_n^{exec}\Big\} + \max\Big\{T_e^{n,off}, T_e^{n,exec}\Big\} \leq T_n, \forall n \in \mathcal{N} \tag{35c}$$

$$E_n^{off} + E_n^{exec} \leq E_n, \forall n \in \mathcal{N} \tag{35d}$$

$$\sum_{n=1}^{N}\Big(E_e^{n,off} + E_e^{n,exec}\Big) \leq E_e. \tag{35e}$$

Regarding the physical meaning and interpretation of the edge server's utility function in Eq. (35a), the term $1 - e^{\frac{2d_n}{D_n}}$ constitutes a strictly increasing and concave function with respect to each user's n amount of offloaded bytes d_n, expressing the edge server's satisfaction, which saturates as its computational burden increases. The remainder of Eq. (35a) constitutes the end-to-end edge-fog computing environment's total energy consumption (overhead), while $\mathcal{C} \in \mathbb{R}^+$ is a cost-of-energy constant factor measured in [1/J]. Concerning the optimization problem's constraints, Eq. (35b) indicates the feasible range of values of each user's amount of offloaded bytes d_n. Eq. (35c) represents each user's n end-to-end completion time requirement, which is calculated as the sum of the maximum time overheads from the transmission and the execution at the user and the edge server layers, assuming that the wireless transmission and computation processes can be performed concurrently. Last, Eq. (35d) and Eq. (35e) guarantee each user device's and the edge server's energy consumption constraint, respectively, where E_e [J] is the edge server's maximum energy constraint.

The optimization problem in Eq. (35a)-(35e) is concave, since the utility function is a sum of concave functions on $d_n, \forall n \in \mathcal{N}$ and the constraints form a compact, i.e., closed and bounded, and convex set. Therefore, in order to derive the optimal solution, which is the vector of optimal amount of offloaded bytes $\mathbf{d}^* = [d_1^*, \ldots, d_n^*, \ldots, d_N^*]$, existing concave/convex optimization tools can be utilized [36].

B. FOLLOWERS' OPTIMIZATION

Given the amount of offloaded bytes d_n for each user n, after broadcasting by the edge server to the users, the users' uplink transmission power control takes place. Specifically, the aim of each user is to distributively maximize its personal transmission-based energy efficiency, by optimizing its uplink transmission power to the edge server, while satisfying its personal transmission time requirement. As a result, the optimization problem to be solved by each user n is given by:

$$\max_{p_n} \; EE_n(p_n, \mathbf{p}_{-n}) = \frac{R_n}{p_n}, \forall n \in \mathcal{N} \tag{36a}$$

$$\text{s.t. } 0 \leq p_n \leq p_n^{max}, \forall n \in \mathcal{N} \tag{36b}$$

$$G_n p_n - \sum_{n'=1}^{n-1} G_{n'} p_{n'} \geq p_{tol}, n = 2, \ldots, N \tag{36c}$$

$$T_n^{off} \leq T_n^{off,max}, \forall n \in \mathcal{N}. \tag{36d}$$

In the above optimization problem, Eq. (36a) represents the user's n energy efficiency utility function, where $\mathbf{p}_{-n} = [p_1, \ldots, p_{n-1}, p_{n+1}, \ldots, p_N]$ is the vector of uplink transmission powers of all users except for user n. The constraint in Eq. (36b) guarantees the user's satisfaction of its maximum uplink transmission power budget p_n^{max}, while Eq. (36c) guarantees the successful decoding of the user's signal via the SIC technique at the edge server's receiver, according to the receiver's sensitivity/tolerance p_{tol}. Eq. (36d) refers to the user's personal transmission time constraint $T_n^{off,max}$ [s].

To capture the interplay among the different users' power control procedure, a non-cooperative game is formulated among them, denoted as $\Pi = [\mathcal{N}, \{\Sigma_n\}_{\forall n \in \mathcal{N}}, \{EE_n\}_{\forall n \in \mathcal{N}}]$, where $\mathcal{N}$ is the set of players, i.e., the users, Σ_n is each user's strategy set of feasible power levels, as imposed by the constraints in Eq. (36b)-(36d), and EE_n is each user's utility function. The non-cooperative game Π is treated as a distributed utility maximization problem, in which each user n updates its uplink transmission power p_n autonomously, by possessing prior information about the other users' transmission power levels $\mathbf{p}_{-n}$, as broadcasted by the edge server.

Towards solving the non-cooperative game Π, the concept of Nash equilibrium is adopted, and the optimal users' strategy vector $\mathbf{p}^* = [p_1^*, \ldots, p_n^*, \ldots, p_N^*]$, from which no user has the incentive to deviate given the strategies of the rest of the users, is determined via a Best Response Dynamics (BRD) algorithm. The interested reader may refer to [40] regarding the definition of the Nash equilibrium, as well as the description of the BRD algorithm. To ensure the existence of at least one Nash equilibrium point for the non-cooperative game Π and thus, the convergence of the users' strategies to the Nash equilibrium, we adopt the theory of the n-person generalized concave games [41].

Theorem 3 (Existence of Nash Equilibrium): The non-cooperative game Π is a n-person generalized concave game and admits at least one Nash equilibrium point, if the following conditions hold true [41]:

1) the strategy sets $\Sigma_1, \ldots, \Sigma_N$ are non-empty, compact, convex subsets of finite dimensional Euclidean spaces,
2) all utility functions $EE_1, \ldots, EE_N$ are continuous on $\Sigma = \Sigma_1 \times \cdots \times \Sigma_N$,
3) every utility EE_n is a quasi-concave function of p_n over Σ_n if all the other strategies are held fixed.

The energy efficiency problem under consideration has been extensively studied in the literature and it is well

Algorithm 1 Incentive and Resource Allocation Framework

Initialize the discrete sets $\mathcal{D}$, Φ, $\mathcal{T}$, B, Γ, Δ.
Calculate the unwillingness-to-participate user types $Z_i, 1 \leq i \leq KLM$ and sort them in ascending order.
Initialize ξ, c, λ, $r(\cdot)$, $q(\cdot)$, $h(\cdot)$.
Design the optimal contract bundles $w_i^* = \{x_i^*, r_i^*\}, 1 \leq i \leq KML$ by solving Eq. (33).
Initialize the set $\mathcal{N}$ and the edge-fog computing environment, including users', edge and fog servers' locations.
Initialize D_n, ϕ_n, T_n, E_n, G_n, p_n^{max}, $T_n^{off,max}$, F_n, σ_n, E_e, G_e, p_e, W_e, a_e, F_e, σ_e, W_f, a_f, ρ, I_0, p_{tol}, $\mathcal{C}$.
Map each user to its optimal contract bundle $w_n^* = \{x_n^*, r_n^*\} \leftrightarrow w_i^* \in W = \{w_i^*, 1 \leq i \leq KLM\}$.
Sort the users in ascending order according to G_n.
Initialize randomly $p_n \in [0, p_n^{max}], \forall n \in \mathcal{N}$.
Set $i = 0$.
repeat
  Set $i = i + 1$.
  Determine the optimal amount of offloaded bytes $d_n^{*(i)}, \forall n \in \mathcal{N}$ by solving Eq. (35a)-(35e).
  Set $j = 0$.
  repeat
    Set $j = j + 1$.
    for $n \in \mathcal{N}$ **do**
      Determine the optimal uplink transmission power $p_n^{*(j)}$ by solving Eq. (36a)-(36d).
    end for
  until $|p_n^{*(j)} - p_n^{*(j-1)}| \leq \epsilon, \forall n \in \mathcal{N}$, where $\epsilon \approx 10^{-5}$.
until $|d_n^{*(i)} - d_n^{*(i-1)}| \leq \epsilon, \forall n \in \mathcal{N}$, where $\epsilon \approx 10^{-5}$.

known that bears the properties that are summarized in Theorem 3 [42], [43], while an extensive proof of Theorem 3 can be found in our prior work in [44]. Given that each follower's, i.e., user's, distributed optimization problem in Eq. (36a)-(36d) is quasi-concave due to the quasi-concave energy efficiency function, this can be effectively treated by applying the Dinkelbach's algorithm [43], [45], which transforms the quasi-concave problem into a series of concave problems that are iteratively solved until convergence. Accordingly, each concave problem can be solved based on existing optimization tools [36].

After the convergence of the non-cooperative game, the users' optimal uplink transmission powers $\mathbf{p}^*$ are fed back to the edge server and the next iteration of the Stackelberg game is established. This procedure is repeated until convergence of the overall Stackelberg game is reached and the Stackelberg equilibrium point $(\mathbf{d}^*, \mathbf{p}^*)$ is found, according to which neither the edge server nor the users have any incentive to deviate from, as shown in Algorithm 1.

C. COMPUTATION COMPLEXITY

To facilitate the derivation of the overall proposed incentive mechanism and resource allocation algorithm's computation complexity, as presented in Algorithm 1, the following algorithmic complexities are considered alone. First, the asymptotic complexity of a convex optimization problem is polynomial in the number of the optimization variables [43], [46], which applies for the optimization problems in Eq. (33), Eq. (35a)-(35e) and Eq. (36a)-(36d), while the Dinkelbach's algorithm has super-linear convergence rate [45], [47]. The users' sorting with respect to their channel gain can be performed with $\mathcal{O}(N^2)$ complexity via the Quicksort algorithm, while for the mapping of each user to its optimal contract bundle a searching algorithm can be employed, e.g., the Binary Search Algorithm, whose algorithmic complexity is $\mathcal{O}(log(KLM))$ in our case, due to the KLM existing contract bundles. The rest typical mathematical manipulations are of $\mathcal{O}(1)$ complexity.

We indicate as I and J the total number of iterations required for the Stackelberg and the non-cooperative game among the users to converge, respectively. Also, following commonly used practises, we denote as I_D the total number of iterations required for the Dinkelbach algorithm to converge, when solving a single user's optimization problem in Eq. (36a)-(36d). As a result, considering that the distributed non-cooperative game among the users is performed in parallel, the overall computation complexity of Algorithm 1 is calculated as $\mathcal{O}(2{\cdot}KLM{+}N{\cdot}log(KLM){+}N^2{+}I{\cdot}(N{+}J{\cdot}I_D{\cdot}1))$. Indicative numerical results that depict the actual number of Stackelberg game iterations, which are required until convergence is met, are enclosed in Section V-B below.

V. EVALUATION AND RESULTS

This section is devoted to the performance evaluation of the proposed incentive mechanism and the joint communication and computing resource allocation procedure, via modeling and simulation. First, we examine the operational characteristics of the multi-dimensional contract-based incentive mechanism, considering, also, the benchmark contract under complete information (as described in Section III-E). Subsequently, we focus on validating the operation and performance of the Stackelberg game-based joint communication and computing resource allocation, accounting for its convergence behavior, as well as comparing it against various alternative baseline offloading approaches. It should be noted that NOMA has been adopted as an underlying technique to facilitate the users' multiplexing and transmissions to the edge.

The simulation setting and the parameters that were used throughout the numerical evaluation that is enclosed in the remainder of this section are initialized as follows. We consider a two-layer edge-fog computing environment, consisting of an edge server, which lies 200 m away from a fog server, and N users deployed with 20-meter increasing distance from the edge server that form a NOMA cluster. Each user has a computing application A_n, whose characteristics can be derived from the following sets: $\mathcal{D} = \{1, 1.2, 1.4\}$ Mbits, $\Phi = \{20, 30, 40\}$ CPU cycles/bit, $\mathcal{T} = \{0.08, 0.1, 0.12\}$ s and $\mathcal{E} = \{1\}$ J. As a result, we

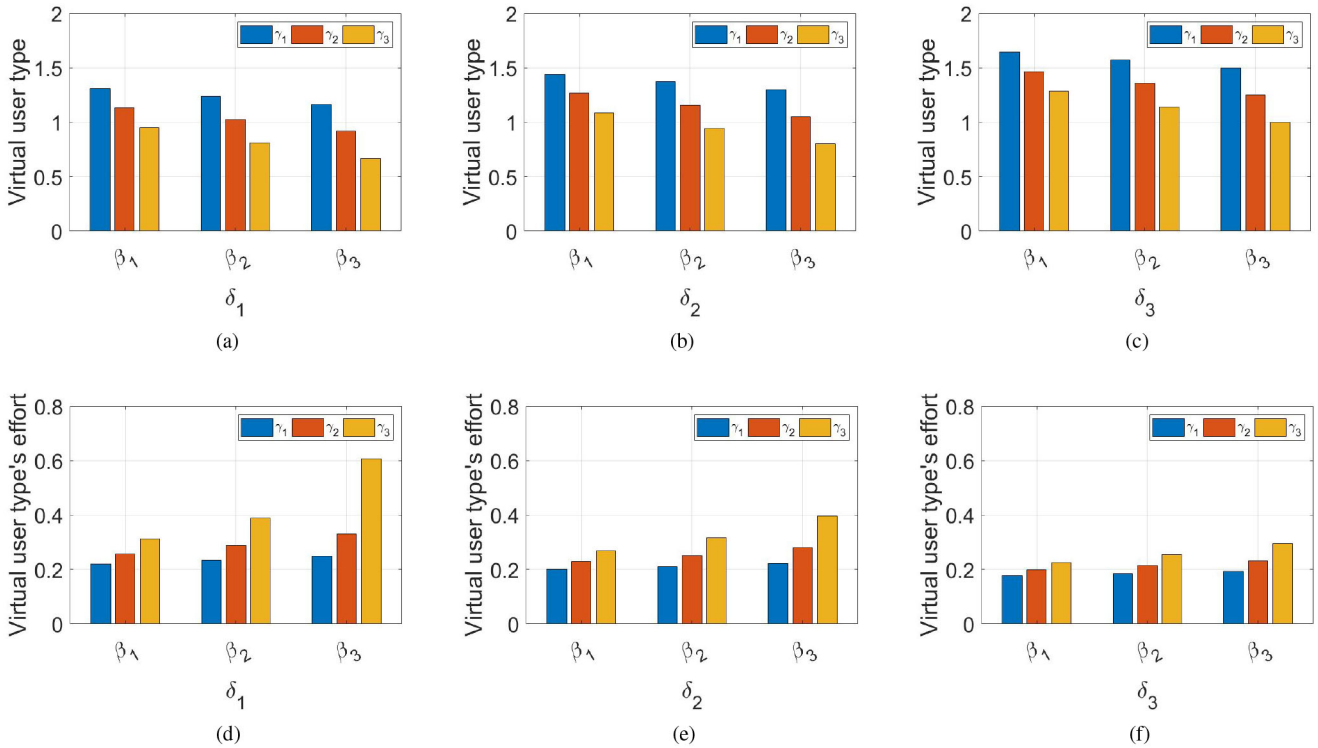

FIGURE 3. Pure evaluation of multi-dimensional contract under different values of the three-dimensional user types (β, γ, δ).

assume that there exist $K \times L \times M = 3 \times 3 \times 3$ different combinations of user application characteristics. The users offload part of their computation tasks at the edge server, while a part of them can be further forwarded by the edge server to the fog, if beneficial. Both the users-to-edge server and edge-to-fog server transmissions are performed wirelessly, while the corresponding bandwidth in the two transmission levels is defined as $W_e = 5$ MHz and $W_f = 1$ MHz, accordingly. Other communications-related parameters are set as: $p_n^{max} = 24$ dBm, $T_n^{off,max} = 0.05$ s, $p_e = 24$ dBm, $a_e = 3.5$, $a_f = 2$, $E_e = 200$ J, $\rho = -20$ dB, $I_0 = -174$ dBm/Hz, $p_{tol} =$ dBm [44]. Considering the computing-related parameters, we consider $F_n = 10^9$ CPU cycle/s and $\sigma_n = 10^{-27}$ J/CPU cycle for each user, and $F_e = 5 \times 10^{11}$ CPU cycles/s and $\sigma_e = 10^{-29}$ J/CPU cycle for the edge server [24]. Last, regarding the multi-dimensional contract we define the parameters $c = 0.5$, $\lambda = 0.8$, $\xi = 1$, while we set $\mathcal{C} = 10^4$ in the Stackelberg game, subsequently.

Finally, for statistical purposes, the numerical results enclosed in Section V-B, below, which pertain to the Stackelberg game-based resource allocation procedure, have been averaged over 100 repetition corresponding to different users' computing application characteristics.

A. EVALUATION OF MULTI-DIMENSIONAL CONTRACT-BASED INCENTIVE MECHANISM

Given that there exist $K \times L \times M = 3 \times 3 \times 3$ different combinations of user application characteristics, then, $3 \times 3 \times 3$ different unwillingness-to-participate user types are formed, capturing the (un)willingness to allow part of their initially offloaded tasks at the edge to be further forwarded and processed at the fog. In this section, we first scrutinize the pure operation of the designed multi-dimensional contract, by analyzing the behavior and trend of the resulting $3 \times 3 \times 3$ unwillingness-to-participate (or virtual) user types and their suited optimal efforts, under different values of the three-dimensional private information (β, γ, δ). In particular, in Fig. 3, the values of the virtual user types and their optimal efforts are depicted as a function of the different values of (β, γ, δ), assuming that are sorted in ascending order as $\beta_1 \leq \beta_2 \leq \beta_3$, $\gamma_1 \leq \gamma_2 \leq \gamma_3$ and $\delta_1 \leq \delta_2 \leq \delta_3$. From Fig. 3(a)-3(c), we observe that as the delay sensitivity evaluation type δ increases, then the values of the virtual user types increase, resulting in lower provided efforts to the edge server in Fig. 3(d)-3(f), which in turn, verify the monotonicity condition of the contract in Lemma 3. Focusing on a single value of δ, e.g., δ_1 in Fig. 3(a), then it can be easily deduced that a low value of either parameter β or γ results in a higher value of the virtual user type, according to the definition of the unwillingness-to-participate parameter in Eq. (17). As a result of the higher unwillingness to participate in the contract, the users' efforts decrease, as shown in Fig. 3(d). The same observation holds for the instances δ_2 and δ_3 regarding the values of the virtual user types in Fig. 3(b)-3(c) and their optimal efforts in Fig. 3(e)-3(f), respectively. This is quite intuitive, since β and γ represent an evaluation of the application size and intensity, higher values of which show

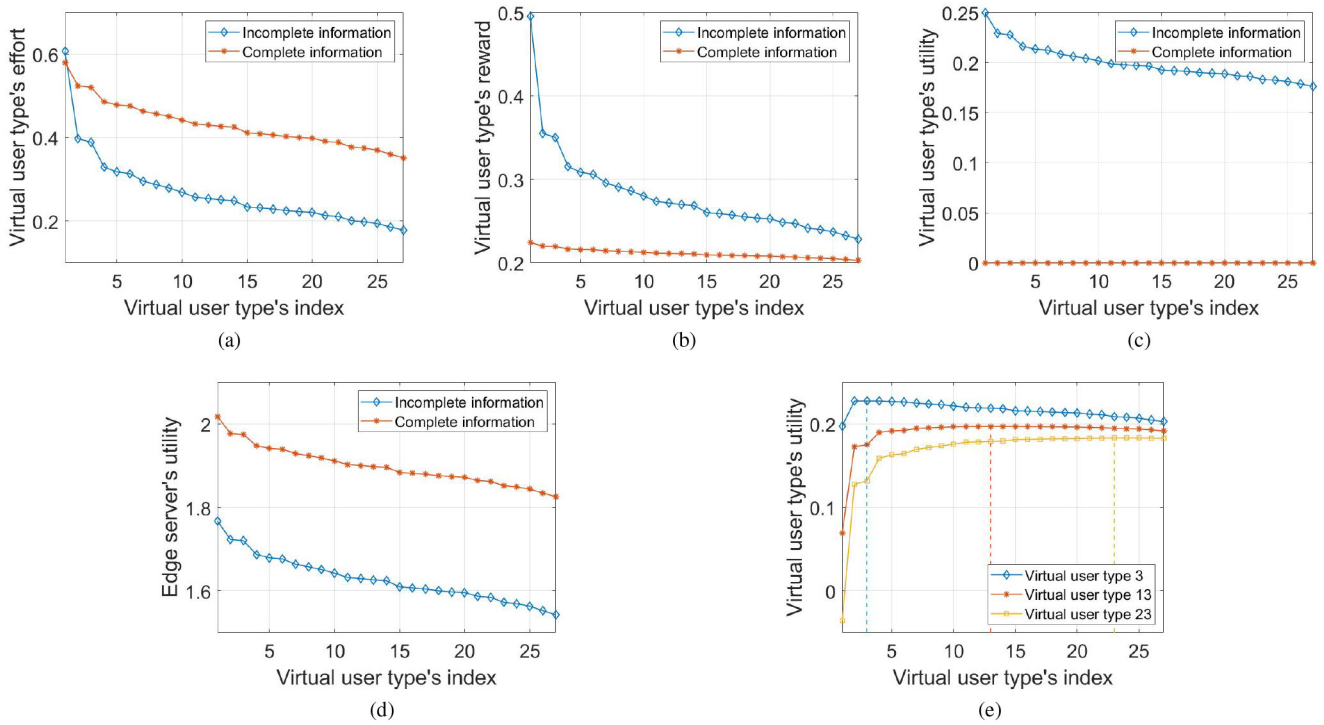

FIGURE 4. Comparative evaluation of multi-dimensional contract under incomplete and complete information cases.

the "need" to participate in the contract in order to offload as many bytes as possible.

Next, we proceed to the comparative evaluation of the proposed multi-dimensional contract, taking into account the benchmark complete information contract case. To this end, in Fig. 4, we consider the $K \times L \times M$ virtual user types sorted in ascending order as $Z_1(x) \leq \cdots \leq Z_i(x) \leq \cdots \leq Z_{KLM}(x)$, $1 \leq i \leq KLM$, indicating them by their sorted index (horizontal axis), and we examine the values of different metrics, such as their efforts, rewards or utilities (vertical axis), under both the incomplete and complete information cases. All graphs in Fig. 4(a)-4(d) validate the monotonic behavior of the designed contract, according to which a higher unwillingness-to-participate/virtual user type provides a lower effort to the edge server, and hence, is rewarded less, yielding at lower utilities for both itself and the edge server. Evidently, in the complete information case, the edge server designs contract bundles that require higher efforts to be provided by the users in exchange for lower rewards compared to the incomplete information case. This naturally stems from the fact that the edge server knows a priori the users' types and fully exploits their efforts, by marginally ensuring their participation in the contract, i.e., the satisfaction of their Individual Rationality (IR) conditions, as expressed in Eq. (34b).

Accordingly, in the complete information case, each virtual user type perceives a zero utility, as illustrated in Fig. 4(c), while the edge server achieves a higher utility per user type under such an ideal complete information availability case compared to the incomplete information one (Fig. 4(d)). In order to complement our evaluation of the multi-dimensional contract-based incentive mechanism, we investigate the derived optimal contract bundles' compliance to the Incentive Compatibility (IC) condition in Definition 2. For this reason, the virtual user types of index 3, 13 and 23 are indicatively selected and their utility values are plotted over all the $K \times L \times M$ contract bundles that have been designed by the edge server (horizontal axis), as shown in Fig. 4(e). Indeed, it can be easily observed that the utility of either virtual user type from the 3, 13 and 23 is maximized when selecting the contract bundle that is tailored to this specific type, verifying the incentive compatibility of the designed contract.

B. EVALUATION OF NON-COOPERATIVE GAME-BASED OFFLOADING MECHANISM

In this section, we aim to elucidate the operational characteristics of the proposed Stackelberg game-based overall communication and computing resource allocation procedure. To this end, we initially study the pure performance and convergence behavior of the proposed Stackelberg game, by examining the progression in the values of different metrics as a function of the different iterations that are required for the game to converge. In particular, Fig. 5(a)-5(b) presents the mean users' transmission power levels and amount of offloaded bytes, accordingly, with respect to the corresponding Stackelberg game iteration index. The different curves that are incorporated in Fig. 5(a)-5(b), correspond to different scenarios with respect to the number of users existing in the system, i.e., $N = \{3, 5, 7, 9\}$, that share the

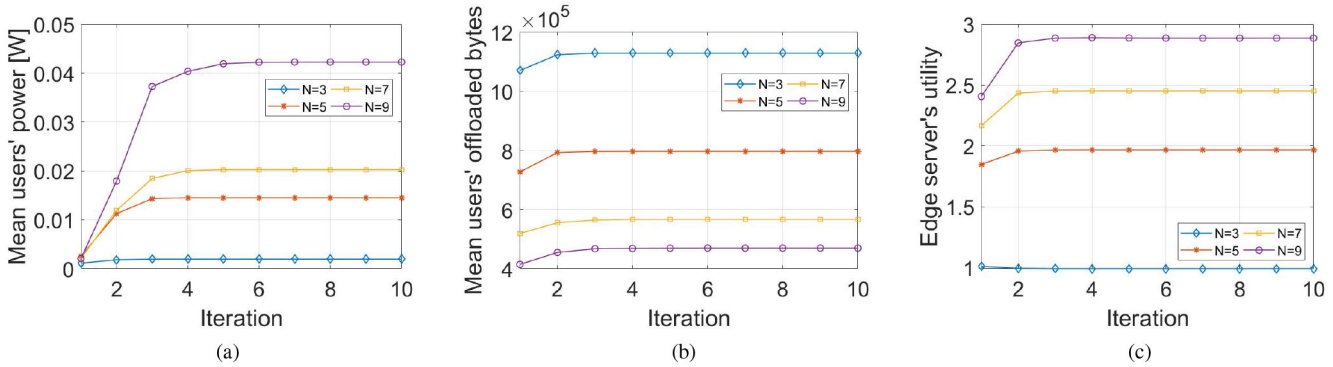

FIGURE 5. Convergence evaluation of Stackelberg game-based resource allocation under different number of users N.

same wireless access bandwidth and are multiplexed via the NOMA technique. Additionally, Fig. 5(c) depicts the edge server's utility, as defined in Eq. (35a) in the leader's optimization problem in Section IV-A, as a function of the Stackelberg game iteration index. The results reveal that the overall interaction between the leader and the followers, via the Stackelberg game, is completed after a small number of iterations (i.e., approximately $I = 6$ iterations for practical purposes), while the number of iterations required increases with the number of the users existing in the NOMA cluster. This can be easier noticed and verified by comparing the curves that regard $N = 3$ and $N = 9$ number of users in Fig. 5(a)-5(c). Furthermore, it is confirmed that as the number of users increases, then their mean consumed uplink transmission power, as derived from the non-cooperative game among them in Section IV-B, increases as well, whereas their mean offloaded bytes to the edge server decrease, as calculated by the leader's optimization problem in Section IV-A. On the one hand, the latter originates from the fact that the overall sensed interference by the users within the NOMA cluster increases, affecting, i.e., reducing, their achieved data rate and hence, increasing the required time to transmit their data. On the other hand, this behavior is, also, encouraged by the edge server's, i.e., the leader's, utility function, which expresses the edge server's dissatisfaction and disutility from the increase in the amount of offloaded bytes by each user and is denoted by the term $1 - e^{\frac{2d_n}{D_n}}$ in Eq. (35a). On the contrary, considering the absolute increase in the sum users' offloaded bytes due to the increase in the number of users in the system, the edge server's utility increases, as can be seen in Fig. 5(c).

Subsequently, we aim to investigate the effectiveness and efficiency of the proposed Stackelberg game-based resource allocation, by comparing it against various alternative baseline offloading approaches. Specifically, in our comparative analysis we consider the cases that the users' tasks are computed: exclusively locally ("Only local"), at the edge ("Only edge") or at the fog ("Only fog"), as well as indicative intermediate cases that 20%, 33% or 40% of the users' total bytes are computed locally, denoted as "20% local", "33% local" and "40% local", respectively, are also taken into account. In these latter cases the rest offloaded amount of bytes, i.e., 80%, 67% and 60% of the overall user application bytes, is equally split between the edge and fog server layers. Last, we, also, invoke the "Random" offloading baseline case, under which a random amount of bytes is offloaded at the edge and fog server layers. At this point, it should be noted that for fairness purposes in all of the aforementioned offloading approaches that require users-to-edge server wireless transmissions, the non-cooperative game among the users that determines their optimal uplink transmission powers to the edge, is performed without exception.

Fig. 6(a) presents the sum users' end-to-end time overhead, which is calculated based on Eq. (35c), with respect to the different offloading approaches and different number of users existing in the system. Apparently, our proposed approach exhibits the lowest sum users' end-to-end time overhead, with the lowest marginal increase with the number of the users, except for the "Only local" case, which is benefited in terms of time by the zero users-to-edge server and edge-to-fog server wireless transmissions. However, as clearly shown in Fig. 6(b), this latter behavior of the "Only local" case occurs at the cost of much higher energy consumption (i.e., worst performance among all alternatives) for the energy-constrained user devices due to local execution, as is discussed later in this section.

In order to better visualize the effect of the increased users' end-to-end time overhead on the satisfaction of their completion time requirement $T_n, \forall n \in \mathcal{N}$, we summarize the percentages of the users that successfully met their time constraints in Table 1, considering all the aforementioned offloading alternatives for different number of users in the system. It should be noted that in the designed simulation setup, the "Only local" case provides always a feasible solution in terms of user time satisfaction, though at the cost of high energy consumption from the users' perspective. As clearly demonstrated by the provided results in Table 1, the proposed approach proves to be the only one among the rest of the examined offloading alternatives that allows for the satisfaction of the users' completion time constraints under varying number of users. Especially, as the number of users

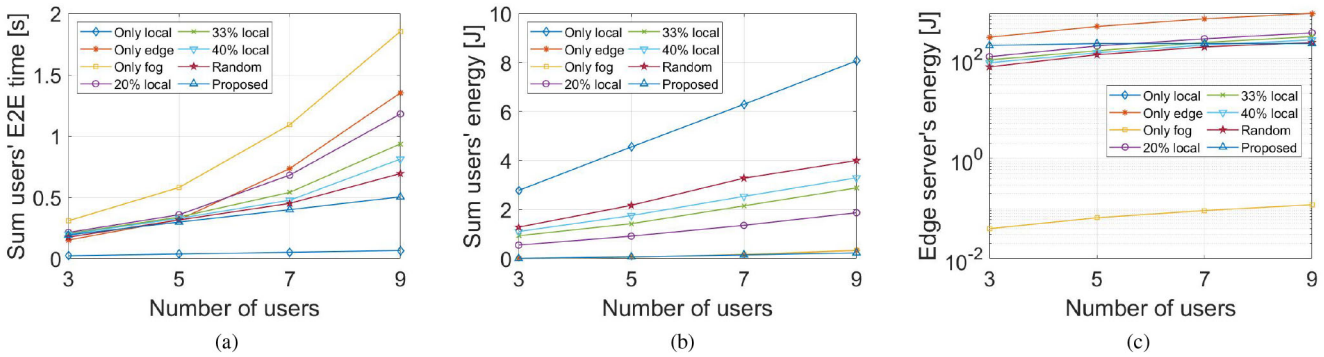

FIGURE 6. Comparative evaluation of Stackelberg game-based resource allocation under different offloading approaches and number of users N.

increases, e.g., $N = 9$, the majority of the baseline offloading approaches fail to satisfy the users' requirements, while in most cases the percentage of satisfied users drops significantly below 50%. Therefore, it becomes apparent, that the dynamic and adaptive features of our proposed optimization-driven approach, achieve the 100% assurance of each user's end-to-end completion time requirement in all considered settings.

Continuing, in Fig. 6(b), the sum users' energy overhead, which accounts for both the wireless transmissions to the edge and their local computation energy consumption, according to Eq. (35d), is depicted, verifying that the proposed approach yields the lowest sum users' energy consumption along with the "Only edge" and "Only fog" baseline cases. However, it is reminded that these other two alternatives, as shown in Table 1, fail to meet the user time constraints for most of the users in most of the different number of users cases, thus resulting in lower user satisfaction percentages compared to the proposed one. Concerning the edge server's energy consumption, which is derived following Eq. (35e) and accounts, also, for both the wireless transmission and the computation energy overheads, Fig. 6(c) demonstrates that in our proposed approach, the edge server operates at its maximum allowed energy consumption point, as denoted by the E_e constraint, while the "Only fog" case intuitively incurs significantly low energy consumption to the edge, at the excessive cost of increased users' energy consumption. It is noteworthy that in Fig. 6(c), the "Only local" case that results in zero edge server's energy consumption is omitted due to the inherent limitation of the logarithmic scale used. Regarding the rest of the alternative scenarios, i.e., "20% local", "33% local", "40% local" and "Random", they appear to yield similar edge server energy consumption, slightly exceeding the edge server's upper energy consumption point E_e as the number of users increases, in contrast to the "Only edge" scenario, which steadily exceeds the edge server's upper energy consumption bound, having to deal with the whole system's computation burden.

TABLE 1. Percentage of users that satisfy their end-to-end time requirement under different offloading approaches.

Approach	$N = 3$	$N = 5$	$N = 7$	$N = 9$
Only local	100%	100%	100%	100%
Only edge	100%	90.8%	50%	34.89%
Only fog	44.67%	32%	15.43%	10.67%
20% local	100%	98.8%	61.14%	36.45%
33% local	100%	100%	86.29%	53.33%
40% local	100%	100%	97.43%	66.89%
Random	97.33%	95.6%	93.71%	81.11%
Proposed	100%	100%	100%	100%

VI. CONCLUSION AND FUTURE WORK

In this paper, the paradigm of edge-fog collaborative computing was promoted, by shifting the selfish users' preference from the prevailing edge service layer to the upper fog computing layer, while accounting for the users' level of delay tolerance. To achieve this, an incentive mechanism was designed, based on a multi-dimensional contract theory model, which allows for the users' characterization and representation by multi-dimensional user types. According to their multi-dimensional types, the users autonomously select their efforts to the edge server, which are mapped to the users' optimal amounts (portions) of their tasks that are initially offloaded at the edge, which can be further forwarded and processed at the fog, in exchange to some reward.

The proposed incentive mechanism was complemented by the users' joint computation task offloading and uplink transmission power allocation via a Stackelberg game played between the edge server and the users. The edge server, having prior information about each user's optimal amount of task that is allowed to be further offloaded at the fog decides the corresponding user's optimal task offloading strategy at the edge, which maximizes its own utility. Next, the users obtain their optimal uplink transmission power levels that maximize their personal communications-related energy efficiency, by participating in a non-cooperative game among themselves. The performance and efficiency of the overall incentive mechanism and resource allocation framework was validated via modeling and simulation, and its superiority and tradeoffs against alternative offloading strategies were demonstrated.

Part of our current and future work refers to the inclusion of multiple edge servers within the two-layer computing environment, which may bear different characteristics and capabilities from both the communication and the computing perspective. Under such a multi-server setting, the problem of the most beneficial and optimal user-to-edge server association needs to be addressed, while at the same time promoting the competition between different edge and/or fog resource providers. In such a multi-server environment, the different edge servers should be inevitably distinguished, not only based on their different communication and computing characteristics, but also based on their offered contract-theoretic rewards back to the users.

REFERENCES

[1] Q.-V. Pham *et al.*, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[2] I. Martinez, A. S. Hafid, and A. Jarray, "Design, resource management, and evaluation of fog computing systems: A survey," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2494–2516, Feb. 2021.

[3] Z. Chen, Q. Ma, L. Gao, and X. Chen, "Edgeconomics: Price competition and selfish computation offloading in multi-server edge computing networks," in *Proc. 19th Int. Symp. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt)*, Philadelphia, PA, USA, 2021, pp. 1–8.

[4] P. Bolton and M. Dewatripont, *Contract Theory*. Cambridge, MA, USA: MIT Press, 2005.

[5] M. Mukherjee *et al.*, "Task data offloading and resource allocation in fog computing with multi-task delay guarantee," *IEEE Access*, vol. 7, pp. 152911–152918, 2019.

[6] H. Shah-Mansouri and V. W. S. Wong, "Hierarchical fog-cloud computing for IoT systems: A computation offloading game," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 3246–3257, Aug. 2018.

[7] Y. Wang, X. Tao, X. Zhang, P. Zhang, and Y. T. Hou, "Cooperative task offloading in three-tier mobile computing networks: An ADMM framework," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2763–2776, Mar. 2019.

[8] E. El Haber, T. M. Nguyen, and C. Assi, "Joint optimization of computational cost and devices energy for task offloading in multi-tier edge-clouds," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3407–3421, May 2019.

[9] L. Li and H. Zhang, "Delay optimization strategy for service cache and task offloading in three-tier architecture mobile edge computing system," *IEEE Access*, vol. 8, pp. 170211–170224, 2020.

[10] W. Lu, S. Hu, X. Liu, C. He, and Y. Gong, "Incentive mechanism based cooperative spectrum sharing for OFDM cognitive IoT network," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 662–672, Apr.–Jun. 2020.

[11] Y. Chen, S. He, F. Hou, Z. Shi, and J. Chen, "An efficient incentive mechanism for device-to-device multicast communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7922–7935, Dec. 2018.

[12] Q. Ma, L. Gao, Y.-F. Liu, and J. Huang, "Incentivizing Wi-Fi network crowdsourcing: A contract theoretic approach," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1035–1048, Jun. 2018.

[13] M. Diamanti, G. Fragkos, E. E. Tsiropoulou, and S. Papavassiliou, "Unified user association and contract-theoretic resource orchestration in NOMA heterogeneous wireless networks," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1485–1502, 2020.

[14] C. Su, F. Ye, T. Liu, Y. Tian, and Z. Han, "Computation offloading in hierarchical multi-access edge computing based on contract theory and Bayesian matching game," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13686–13701, Nov. 2020.

[15] C. Yang, W. Lou, Y. Liu, and S. Xie, "Resource allocation for edge computing-based vehicle platoon on freeway: A contract-optimization approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15988–16000, Dec. 2020.

[16] J. Zhao, M. Kong, Q. Li, and X. Sun, "Contract-based computing resource management via deep reinforcement learning in vehicular fog computing," *IEEE Access*, vol. 8, pp. 3319–3329, 2020.

[17] Y. Li, J. Zhang, X. Gan, L. Fu, H. Yu, and X. Wang, "A contract-based incentive mechanism for delayed traffic offloading in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5314–5327, Aug. 2016.

[18] Z. Xiong, J. Kang, D. Niyato, P. Wang, H. V. Poor, and S. Xie, "A multi-dimensional contract approach for data rewarding in mobile networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5779–5793, Sep. 2020.

[19] W. Y. B. Lim *et al.*, "Towards federated learning in UAV-enabled Internet of Vehicles: A multi-dimensional contract-matching approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5140–5154, Aug. 2021.

[20] Z. Wang, L. Gao, and J. Huang, "Multi-cap optimization for wireless data plans with time flexibility," *IEEE Trans. Mobile Comput.*, vol. 19, no. 9, pp. 2145–2159, Sep. 2020.

[21] G. Mitsis, E. E. Tsiropoulou, and S. Papavassiliou, "Data offloading in UAV-assisted multi-access edge computing systems: A resource-based pricing and user risk-awareness approach," *Sensors*, vol. 20, no. 8, p. 2434, 2020.

[22] P. A. Apostolopoulos, E. E. Tsiropoulou, and S. Papavassiliou, "Risk-aware data offloading in multi-server multi-access edge computing environment," *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1405–1418, Jun. 2020.

[23] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal resource allocation for delay minimization in NOMA-MEC networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7867–7881, Dec. 2020.

[24] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 310–313, Feb. 2019.

[25] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in MEC systems with NOMA," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2191–2206, Apr. 2021.

[26] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. New York, NY, USA: Cambridge Univ. Press, 2011.

[27] R. Mahmud, R. Kotagiri, and R. Buyya, "Fog computing: A taxonomy, survey and future directions," in *Internet of Everything: Algorithms, Methodologies, Technologies and Perspectives*, B. Di Martino, K.-C. Li, L. T. Yang, and A. Esposito, Eds. Singapore: Springer, 2018, pp. 103–130.

[28] S. S. Gill, "A manifesto for modern fog and edge computing: Vision, new paradigms, opportunities, and future directions," in *Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases*, R. Nagarajan, P. Raj, and R. Thirunavukarasu, Eds. Cham, Switzerland: Springer Int., 2022, pp. 237–253.

[29] M. Goudarzi, H. Wu, M. Palaniswami, and R. Buyya, "An application placement technique for concurrent IoT applications in edge and fog computing environments," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1298–1311, Apr. 2021.

[30] M. Heck, J. Edinger, D. Schaefer, and C. Becker, "IoT applications in fog and edge computing: Where are we and where are we going?" in *Proc. 27th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Hangzhou, China, 2018, pp. 1–6.

[31] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[32] S. Mao, J. Wu, L. Liu, D. Lan, and A. Taherkordi, "Energy-efficient cooperative communication and computation for wireless powered mobile-edge computing," *IEEE Syst. J.*, early access, Oct. 15, 2020, doi: 10.1109/JSYST.2020.3020474.

[33] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Distributed resource allocation and computation offloading in fog and cloud networks with non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12137–12151, Dec. 2018.

[34] P. A. Apostolopoulos, G. Fragkos, E. E. Tsiropoulou, and S. Papavassiliou, "Data offloading in UAV-assisted multi-access edge computing systems under resource uncertainty," *IEEE Trans. Mobile Comput.*, early access, Mar. 31, 2021, doi: 10.1109/TMC.2021.3069911.

[35] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.
[36] T. Coleman, M. A. Branch, and A. Grace, *Optimization Toolbox For Use with MATLAB: User's Guide, Version 2, Release II*, MATLAB, Natick, MA, USA, 1999.
[37] G. Zheng, C. Xu, H. Long, and X. Zhao, "MEC in NOMA-HetNets: A joint task offloading and resource allocation approach," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Nanjing, China, 2021, pp. 1–6.
[38] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, Apr. 2021.
[39] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Advances in Soft Computing (AFSS)*, N. R. Pal and M. Sugeno, Eds. Berlin, Germany: Springer, 2002, pp. 288–300.
[40] E. E. Tsiropoulou, G. K. Katsinis, and S. Papavassiliou, "Distributed uplink power control in multiservice wireless networks via a game theoretic approach with convex pricing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 1, pp. 61–68, Jan. 2012.
[41] F. Forgó, "On the existence of Nash-equilibrium in *n*-Person generalized concave game," in *Generalized Convexity*, S. Komlósi, T. Rapcsák, and S. Schaible, Eds. Berlin, Germany: Springer, 1994, vol. 405, pp. 53–61.
[42] H. Yu, Y. Li, M. Kountouris, X. Xu, and J. Wang, "Energy efficiency analysis of relay-assisted cellular networks," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–11, 2014.
[43] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.
[44] M. Diamanti, P. Charatsaris, E. E. Tsiropoulou, and S. Papavassiliou, "The prospect of reconfigurable intelligent surfaces in integrated access and backhaul networks," *IEEE Trans. Green Commun. Netw.*, early access, Nov. 15, 2021, doi: 10.1109/TGCN.2021.3126784.
[45] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Sci.*, vol. 13, no. 7, pp. 492–498, 1967.
[46] J. Xiong, L. You, D. W. K. Ng, C. Yuen, W. Wang, and X. Gao, "Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, 2020, pp. 1–6.
[47] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," in *Foundations and Trends in Communications and Information Theory*, vol. 11. Hanover, MA, USA: Now Publ. Inc., 2015, pp. 185–396.

MARIA DIAMANTI (Graduate Student Member, IEEE) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2018. She is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering, National Technical University of Athens, where she is also a Research Assistant. Her research interests lie in the areas of 5G/6G wireless networks, resource management and optimization, game theory, contract theory, and reinforcement learning.

PANAGIOTIS CHARATSARIS received the Diploma degree in electrical and computer engineering from the National Technical University of Athens in 2021, where he is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering. His overall research interests lie in the broader area of resource optimization in 5G/6G wireless communications systems.

EIRINI ELENI TSIROPOULOU (Senior Member, IEEE) is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of New Mexico. Her main research interests lie in the area of cyber–physical social systems and wireless heterogeneous networks, with emphasis on network modeling and optimization, resource orchestration in interdependent systems, reinforcement learning, game theory, network economics, and Internet of Things. Four of her papers received the Best Paper Award at IEEE WCNC in 2012, ADHOCNETS in 2015, IEEE/IFIP WMNC 2019, and INFOCOM 2019 by the IEEE ComSoc Technical Committee on Communications Systems Integration and Modeling. She was selected by the IEEE Communication Society—N2Women—as one of the top ten Rising Stars of 2017 in the communications and networking field. She received the NSF CRII Award in 2019 and the Early Career Award by the IEEE Communications Society Internet Technical Committee in 2019.

SYMEON PAPAVASSILIOU (Senior Member, IEEE) is currently a Professor with the School of ECE, National Technical University of Athens. From 1995 to 1999, he was a Senior Technical Staff Member with AT&T Laboratories, Middletown, NJ, USA. In August 1999, he joined the ECE Department, New Jersey Institute of Technology, USA, where he was an Associate Professor until 2004. He has an established record of publications in his field of expertise, with more than 350 technical journal and conference published papers. His main research interests lie in the area of computer communication networks, with emphasis on the analysis, optimization, and performance evaluation of mobile and distributed systems, wireless networks, and complex systems. He received the Best Paper Award at IEEE INFOCOM 94, the AT&T Division Recognition and Achievement Award in 1997, the U.S. National Science Foundation Career Award in 2003, the Best Paper Award at IEEE WCNC 2012, the Excellence in Research Grant in Greece in 2012, the Best Paper Awards at ADHOCNETS 2015, ICT 2016, and IEEE/IFIP WMNC 2019, as well as the 2019 IEEE ComSoc Technical Committee on Communications Systems Integration and Modeling Best Paper Award (for his INFOCOM 2019 paper). He also served on the board of the Greek National Regulatory Authority on Telecommunications and Posts from 2006 to 2009.