An Efficient Tasks Offloading Procedure for an Integrated Edge-Computing Architecture

Benedetta Picano and Romano Fantacci

***Abstract*—The advent of sixth-generation networks has given rise to numerous challenges, requiring the synergistic exploitation of both ground and air edge computing facilities. This paper considers an integrated ground-air edge computing scenario where the computation offloading of a set of delay sensitive tasks has to be performed in a context where ground and air computational facilities are already involved in monitoring and control proceduers in a remote area under an unpredictable overload of computation requests, e.g., related to the management of an emergency situation. In this reference, a matching game is proposed to assign tasks to the most suitable computation nodes, in order to minimize the outage probability of the newly arrived tasks, i.e., the probability with which tasks experience a completion time greater than the corresponding deadline. To this regard, we have considered that new allocated task suffer for a waiting time due to the time needed to complete the service of all the tasks already in the ground or air computation node. As a consequence, to statistically characterize such waiting time, under proper assumptions, we have resorted to the G/G/1 queuing system model and the Lindley's integral equation approach to define a suitable metric to formulate a tasks allocation procedure based on the matching theory. Furthermore, matching stability has been theoretically proved for the proposed approach. Finally, numerical results have been provided in order to highlight the better behavior of the proposed task allocation scheme in comparison with different state-of-the-art alternatives.**

***Index Terms*—Queueing system, task offloading, unmanned aerial vehicle.**

I. Introduction

THE upcoming sixth-generation (6G) networks has to support a wide plethora of disruptive applications, typically demanding for high rate, high reliability, low latency, and requiring seamless coverage. In such a context, the exclusive use of classical cellular networks seems to be not enough to handle the huge amount of data expected to be injected in the network by the new generation applications, and simultaneously guaranteeing the corresponding high-flying quality of service (QoS) constraints. Within this challenging context, the exclusive exploitation of terrestrial networks cannot meet the far-reaching traffic demand and service quality level, mainly in terms of delay constraints, of the emerging applications. In particular, it appears to date that it is through a synergism carefully orchestrated between terrestrial and non-terrestrial computational resources, that services delay mitigation can be reached [1].

Manuscript received May 17, 2023; approved for publication December 18, 2023; approved for publication by Choi, Jin-Ghoo Division 3 Editor, January 23, 2024.

The authors are with University of Florence, 50139 Firenze, Italy, email:{benedetta.picano, romano.fantacci}@unifi.it.

R. Fantacci is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2024.000004

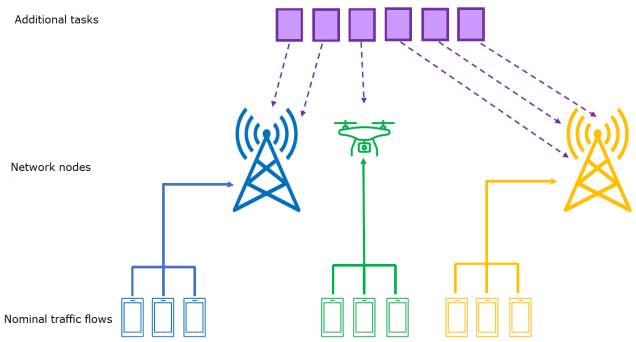

Fig. 1. System scenario where a hybrid ground-aerial domain was considered.

In particular, this offers the promising opportunity to extend and empower the emerging edge computing paradigm [2] in an integrated heterogeneous scenario by considering flying unmanned aerial vehicles (UAVs) able to provide edge computing close to the end-users functionally integrated with ground edge nodes (ENs). This feature is of special interest whenever a quick and temporary computation resource update is needed in contexts where a permanent expansion of the ground ENs infrastructure is not advisable or viable, e.g., performing monitoring and control operations in a remote area under an unpredictable overload of computational demands. More generally, it can be stated that this novel edge computing paradigm aims at overcomes the limits of a classical ground edge computing system in any congested areas, providing faster computation and hosting processing of tasks stemmed from devices in an efficient and flexible manner by enabling tasks offloading towards different ENs locations, e.g., ground or air. In such a context, it becomes of paramount importance to provide a suitable tasks offloading procedure in order to select the most convenient computation site between a set of ground ENs or UAV-ENs. Furthermore, the end-to-end task delay, expressing the time elapsed since a device demanding for task computation submits its request until the device receives the processing outcome, represents a crucial metric to be investigated here. This holds particular significance in multimedia services, ensuring a predefined service quality target. However, it also proves beneficial for critical applications, especially when assessing the risk level associated with potentially surpassing a specific deadline in data delivery, is considered as a key performance indicator.

As a consequence, it has become mandatory to have adequate methodologies to perform a suitable tasks offloading

1229-2370/24/$10.00 © 2024 KICS

and to predict the achievable performance. During years, Markovian analysis has been extensively applied to this regard, providing results in network performance evaluation in an easy way. Nevertheless, in order to provide accurate network performance predictions in the novel application contexts, the assumption of Markovian models may represent a too simplistic hypothesis. In fact, even if a more general system characterization such as G/G/1 model increases the complexity of the theoretical study, the corresponding analysis typically fits better the actual system dynamic, giving rise to a more effective system design [3].

In this reference, the main contributions of the paper can be summarized as follows

- Analysis of the queue behavior at the computation nodes in an integrated gound-air computation system by resorting to the G/G/1 model and use of the Lindley's integral equation, solved by means of spectrum factorization [4] method;
- Design and development of an algorithm to perform tasks offloading with the aim at minimizing the task outage probability, i.e., the probability that a task experiences a completion time greater than the corresponding deadline, based on the queue state conditions of the selected computation nodes, i.e., ground EN or an unmanned aerial vehicles-mounted ENs (UAV-ENs);
- Numerical simulations to test the performance of the framework proposed and validate its effectiveness in comparison with different tasks allocation schemes.

The rest of the paper is organized as follows. In Section II an in-depth review of the related literature is provided. Section III presents both the system model and the problem formulation, whereas Section IV details the proposed framework. Performance evaluations are presented in Section V. Finally, our conclusions are outlined in Section VI.

II. Related Works

Many works are available from the literature regarding the use of UAVs to enable advanced services and applications. In particular, authors in [5] proposed a heuristic to maximize the cellular users coverage optimizing the drones deployment, as long as minimizing the communication cost among UAVs. A drone-as-a-service market model has been developed in [6], in which a service algorithm has been designed, in order to properly meet the quality requirements in terms of cost and delay, expressed by users. Moreover, this paper deals with the improvement of security for delivery drones. In particular, the authors propose a consumer authentication hybrid computing framework for drone delivery as a service, the effectiveness of which is demonstrated through experimental results. In [7] the main focus is the limited UAVs resources. In fact, the authors analyzed the feasibility of overcoming these constraints by combining and controlling multiple UAVs. In this reference, the paper explores programmable crowd-powered drones to create a federated cloud. Moreover, a scripting language is applied to coordinate flight trajectories of multiple drones, as well as multi-drone service management. Differently, in [8], a mixed integer programming problem has been formulated considering the traveling salesman problem with a drone station. The route distortion problem has been defined, and a lower bound of the number of drones needed to solve it has been proposed. Furthermore, the paper [9] proposes an UAV-assisted MEC network with air–ground cooperation, in which both UAV and ground access points exhibit a direct link towards devices and cooperate to execute tasks computation, aiming at minimizing the worst delay and optimizing the resource allocation by jointly controlling UAV-device matching, UAV horizontal and vertical position, bandwidth selection, and task splitting. A two-layered decision-making framework for the cooperation between one or more stations and one or more drones is presented in [10], maximizing profit, and minimizing the travel distance.

G/G/1 queuing systems have been the object of the analysis proposed in [11], in which a shift parameter has been introduced to model the time lag under the assumption of exponential, hyperexponential, and Erlang distributions. Furthermore, paper [12] addresses multiple vehicle-to-vehicle (V2V) connections sharing spectrum with multiple capacity-hungry links. The main goal of the paper has been the resource allocation and packet sampling rate optimization of V2V connections, aiming maximizing the sum ergodic capacity of the network, as well as guaranteeing the age-of-information outage probability of V2V links. Authors in [13] propose an online optimization of both the UAV trajectory and the user association, in order to reach finite queueing delay minimizing UAVs energy consumption. In paper [14], a G/D/1 queuing model for the analysis of network-on-chip has been proposed and studied by resorting to a Jackson queuing network. Authors in [15] consider a hybrid aerial-ground scenario, where the offloading is exploited to perform UAV visual target tracking. Such a deep learning task is sent to an EC node, to meet the constraints on the computational resource and energy capacity typical of UAVs. The aerial-terrestrial communication links are also studied in paper [16], where the multi-task learning is adopted in combination with the reflecting intelligent surfaces (RIS) to improve coverage. An adaptive RIS-assisted transmission protocol, where the channel estimation, the transmission policy, and the data transmission are independently implemented in a frame, is designed. Differently, in [17], a multi-task resource scheduling framework exploiting the deep reinforcement learning has been designed with the objective to minimize the energy consumption of all users and UAVs in the system. Authors in [18] integrate horizontal federated learning with double deep Q-network to solve the problem of the computation offloading and relay communication in air-ground integrated networks, considering emergency scenarios. In reference to the offloading problem, the objective is the minimization of the weighted sum of both delay and energy consumption. For the data transmission, the main goal is to maximize the minimum rate of relay links.

An autonomous network resource demand prediction scheme has been proposed in [19], exploiting queuing theory. More in depth, the analysis addressed in [19], accurately measures the mathematical expectation of queuing length experienced by packets, and the average occupied resources usage in

nodes, taking into account the success rate of communication resource transmission and the limited caching availability on nodes. Then, paper [20] focuses on multipath technology within the context of Internet of things (IoT) monitoring and control applications. Two paths with different strategies have been considering, exploit redundancy and coding to enhance timing performance of wireless communications. In this picture, data blocks have been modeled via a Markovian and a deterministic process, respectively. Then, the packet delay and the peak age of information metrics have been analyzed. While there is a wide range of works that deal with studying computational offloading scenarios where UAVs and edge computing coexist, and other works extensively apply the results of queuing theory, our work combines Lindley's analysis with this kind of landscape. In fact, for the best of authors' knowledge, this is the first paper introducing the Lindley's analysis to obtain a measure of the outage probability within an offloading problem.

III. Problem Statement

A. System Scenario

As depicted in Fig. 1, in this paper, we mainly focus on a scenario where we have an edge computing system devotes to monitoring and control a process in a remote area to face an unpredictable overload of computation requests, e.g., related to the management of an emergency situation. In this case, an efficient and flexible approach is to resort to the deployment of a swarm of UAVs having on board computation capabilities in order to enable an integrated ground-air edge computing infrastructure (ECI), where computation capabilities are located at the ground ENs or on the UAV-ENs. In particular, we have a set $\mathcal{C}$ of processing nodes formed by a set of ground ENs $\mathcal{S} = \{1, \cdots, S\}$ able to perform computation, and a swarm of UAVs $\mathcal{V} = \{1, \cdots, V\}$, each one having on-board computation capabilities and linked to almost one SBS of the ground cellular network. Hence, we have an overall set $\mathcal{C} = \{1, \cdots, S\} \cup \{1, \cdots, V\}$ of possible computation nodes. Note that the elements of both sets $\mathcal{S}$ and $\mathcal{V}$ are heterogeneous in terms of computational capabilities, i.e., mean computation time, with UAV-ENs less powerful in terms of processing speed than the ground ENs.

B. General System Assumptions

To complete the definition of the system scenario, we have made the following assumptions.

- Any device can access the ground-air ECI by means of the most suitable SBS station of cellular network in order to offload its task computation request to any possible EN site belonging to $\mathcal{C}$;
- Any ground EN can be reached by a device through the ground cellular network to offload its task independently from the access SBS, i.e., the SBS directly linked to it;
- Any UAV assigned to the service area of interest can be linked[1], to almost one SBS, i.e., it belongs to $\mathcal{V}$ and, hence, to $\mathcal{C}$, with a given probability p_c;
- All the SBSs of the access cellular network form a *full connected* network, i.e., the task offloading to a given UAV-EN can be routed to the most suitable SBS (if any), linked to the UAV-EN of interest, independently of the access SBS. However, in such a context we have to take into account that each element c in $\mathcal{C}$ is involved in providing computation service, according to a first-in-first-out (FIFO) policy, to previously allocated task flows related to control and monitoring procedures already activated in the operation area. In particular, such task flows (nominal flows in what follows) are characterized by independent general arrival and service processes with mean rate λ_c tasks/s, μ_c tasks/s, respectively, with $\mu_c > \lambda_c$ in order to guarantee the stability at each computation node.

In such a context, we have to provide a suitable computation allocation for an additional set $\mathcal{U} = \{1, \cdots, U\}$ of tasks, related to new needs to make current procedures more specific to the context of interest. Each additional tasks is associated to only one device so that devices and tasks are cited interchangeably in what follows. Furthermore, in our model we have assumed that each task u requires computation completion with a given soft-deadline delay constraint $t_{d,u}$, $u \in \mathcal{U}$. Such a constraint means that tasks are allocate even if their deadline is not satisfied. In fact, soft-deadline applications prefer to receive a delayed service rather than not receiving it at all. Moreover, all tasks in $\mathcal{U}$, allocated to a given node in $\mathcal{C}$, due to the adopted FIFO policy, access the service according to their allocation order and first with respect to any tasks belonging to the nominal flows arrived at that node after their allocation. However, all the nominal flows arrived before and waiting for service in the ground EN or UAV-EN queue maintain the acquired FIFO priority for access the service facility with respect to new computation arrivals.

C. Analytical Approach

The initial step in our analysis aims to determine the waiting time experienced by tasks in the set $\mathcal{U} = \{1, \cdots, U\}$, at the computation nodes due to the workload already allocated (i.e., nominal flows). Hence, our goal is to derive the probability density function (pdf) of the waiting time at a given computation node, modeled as a G/G/1 system, due to the presence of previously arrived tasks belonging to the nominal flow of that node.

In particular, the system under consideration is one where all the possible computation nodes exhibit independent inter-arrival times between tasks belonging to the relative nominal flow, with a general pdf $A_c(t)$, for $\forall c \in \mathcal{C}$. Similarly, the

[1]This simple Bernoulli statistical model takes into account the fact that, due to the UAVs motion, it may not always be possible to have connectivity with at least one ground SBS. For simplicity, without loosing generality of our analysis, we have assumed that the UAVs connection conditions does not change during the completion of the tasks allocation planning and that the UAV connection probability, p_c, with almost one ground SBS is the same for all UAVs.

task computation service times are independent with a general pdf $B_c(t)$, for $\forall c \in \mathcal{C}$. In addition to this, we have assumed that only one server is available at each processing node and that the computation service, as stated before, is performed according to the FIFO policy. Understanding the behavior of a such G/G/1 system, even if under the FIFO queuing policy, typically poses a non-trivial challenge, requiring an appropriate methodology. Among various alternatives, the analytical approach outlined in this paper leads to the *Lindley's integral equation* with solution obtained by resorting to the *spectral factorization method* [4]. Towards this end, by focusing on a given node $c \in \mathcal{C}$, we have that $Q_c^{(n+1)}$ is the waiting time (in queue) experienced by the $(n+1)$th task of the nominal data flow allocated on node c, given by

$$Q_c^{(n+1)} = \begin{cases} Q_c^{(n)} + x_c^{(n)} - h_c^{(n)}, & \text{if } Q_c^{(n)} + x_c^{(n)} - h_c^{(n)} > 0; \\ 0, & \text{if } Q_c^{(n)} + x_c^{(n)} - h_c^{(n)} \leq 0, \end{cases} \tag{1}$$

where $x_c^{(n)}$ is the computation (i.e., service) time of the task $O_c(n)$ belonging to the nominal data flow allocated at node c, whereas $h_c^{(n)}$ represents the interarrival time between two consecutive computation requests, i.e., $O_c(n)$, $O_c(n+1)$ [4]. Denoting with $d_c^{(n)}$ the difference $x_c^{(n)} - h_c^{(n)}$, i.e., $d_c^{(n)} = x_c^{(n)} - h_c^{(n)}$, and considering the stochastic process $\{Q_c^{(n)}, n = 0, 1, \cdots\}$, we have that

$$Q_c^{(n+1)}(t) = Q_c^{(n)} + d_c^{(n)}. \tag{2}$$

Furthermore, being $\mu_c > \lambda_c$ for all $c \in \mathcal{C}$ we have that the stability condition is verified for all the processing nodes, hence, we have :

$$\lim_{n\to\infty} P_r\left[Q^n(c) \leqslant t\right] = W_c(t) \tag{3}$$

with $W_c(t)$ the stationary cumulative probability distribution (CDF) for the waiting time in the queue at node c. Likewise, the stationary CDF for the random variable d_c assumed independent of n, results in :

$$\begin{aligned} C_c^*(u) &= P_r(x_c - h_c \leqslant u) \\ &= \int_{t=0}^{\infty} P_r(x_c \leqslant u + t | h_c = t) A_c(t) dt. \end{aligned} \tag{4}$$

Being, x_c independent of h_c, we have:

$$C_c(u) = \int_{t=0}^{\infty} B_c(u+t) A_c(t) dt, \tag{5}$$

where $C_c(u)$ is the *pdf* of the random variable d_c. Hence, through some algebraic manipulations detailed in [4], skipped here due to space limitation, we obtain the *Lindley's integral equation* in the following form:

$$W_c(t) = \int_{-\infty}^{t} W_c(t-u) C_c(u) du \quad t \geqslant 0, \tag{6}$$

and $W_c(t) = 0$ for $t < 0$.

For the purpose of our analysis, we introduce the function $\phi_{+c}(s)$ defined as:

$$\phi_{+c}(s) = \int_{-\infty}^{+\infty} W_c(t) e^{-st} dt, \tag{7}$$

that can be easily recognized ad the *Laplace* transform of $W_c(t)$. In solving (6) with respect to $W_c(t)$, we resort here to the use of the *spectrum factorization* approach [4]. In particular, being $\mathcal{A}_c(s)$ and $\mathcal{B}_c(s)$ the Laplace transform of $A_c(t)$ and $B_c(t)$, respectively, the goal of the *spectrum factorization* approach, as detailed in [4], is to define two rational functions of s, $\Psi_{+c}(s), \Psi_{-c}(s)$, so that we have:

$$\mathcal{A}_c(-s)\mathcal{B}_c(s) - 1 = \frac{\Psi_{+c}(s)}{\Psi_{-c}(s)}. \tag{8}$$

Moreover, the factorization provided in (8) has to respect the following properties [4]:

- For $Re(s) > 0$, $\Psi_{+c}(s)$ results to be analytic with no zeros in the half-plane;
- $\exists D$ such that for $Re(s) < D$, $\Psi_{-c}(s)$ results to be analytic with no zeros in the half-plane.

Once the factorization is provided, by resorting to the application of the *Liouville's theorem* (see [4] for more details) the Laplace transform of the waiting time CDF, $\phi_{+c}(s)$, results in

$$\phi_{+c}(s) = \frac{K_c}{\Psi_{+c}(s)}. \tag{9}$$

Moreover, being by definition :

$$\lim_{s\to 0} s\phi_{+c}(s) = 1 \tag{10}$$

from (9) we have:

$$K_c = \lim_{s\to 0} \frac{\Psi_{+c}(s)}{s}. \tag{11}$$

Therefore, $W_c(t)$ can be obtained by anti-transforming $\phi_{+c}(s)$ defined in (9) as

$$W_c(t) = \mathcal{L}_s^{-1}[\phi_{+c}(s)](t). \tag{12}$$

Hence, the pdf of the waiting time at node c can be derived as:

$$w_c(t) = \frac{dW_c(t)}{dt}, \tag{13}$$

which completes our analysis concerning the waiting time characterization at each computation node due to the presence of the nominal flows.

D. Problem Formulation

This paper aims at minimizing the outage probability for the additional set of tasks $\mathcal{U}$ needing to be offloaded on the pool of heterogeneous computation nodes $\mathcal{C}$ previously committed to provide computation service to nominal data flows. Towards this end, we have to take into account that the set $\mathcal{S}$ of SBS nodes offer a fast service time with respect to set $\mathcal{V}$ of UAVs alternatives, mainly due to the different energy constraints. In performing our analysis, we have overlooked the set-up time required to acquire exclusive use of a communication channel by users. This assumption is obviously valid in the case of systems with dedicated channels (i.e., safety-critical applications) but it may be considered reasonable also for a random contention set-up phase, in particular, for the case under consideration, where we have a relatively low to moderate number of users linked to a same SBS. In addition, we did

not consider transmission times in defining the deadlines, as according to [21], these times are negligible in our scenario. Nevertheless, as communication times usually represent a constant delay contribution, the proposed analysis can readily extend to encompass scenarios where their impact cannot be deemed negligible by a proper definition of the actual time deadline.

In particular, we propose in the next section an offloading scheme based on a matching algorithm to produce as outcome the allocation matrix $\mathbf{A} \in \{0,1\}^{U\times(S+V)}$, whose generic element $\alpha_{u,c} = 1$ if the task of the generic user u belonging to $\mathcal{U}$ is allocated on the network node c, zero otherwise. The aim of our offloading scheme is to minimizing the outage probability for any task $u \in \mathcal{U}$ whose computation service time[2] $s_{c,u}$ at node c is assumed known. Hence, the outage probability for the generic task u with deadline $t_{d,u}$, offloaded on node c, hence, having computation time $s_{c,u}$ and suffering of a waiting time with pdf $w_c(t)$ due to the nominal flow perviously allocated on c , results in:

$$P_{out,c,u} = \int_{\epsilon_{c,u}}^{\infty} w_c(t)dt = 1 - W_c(\epsilon_{c,u}), \tag{14}$$

where we have:

$$\epsilon_{c,u} = t_{d,u} - s_{c,u} - \sum_{j\in\mathcal{U}\setminus\{u\}} s_{c,j}.\alpha_{j,c}. \tag{15}$$

In this reference, the main objective of the paper is represented by the minimization of the mean outage probability of tasks belonging to $\mathcal{U}$, i.e.,

$$\min_{\mathbf{A}} \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} P_{out,c,u}. \tag{16}$$

In this reference, in the next section a proper tasks allocation procedure on network nodes, heterogeneous in computational capability, is designed, in order to provide a tasks assignment arrangement able to minimize the number of offloaded tasks in outage, on the basis of the queue congestion at all the possible computation nodes.

In summary, the functional requirements of the proposed systems revolve around executing an effective allocation policy to minimize the outage probability of incoming tasks. This objective is achieved by employing a matching algorithm that considers task deadlines, the workload assigned to each computation node, and applies the Lindley's equation. Additionally, the system ensures FIFO priority for tasks previously accepted by computation nodes, while newly allocated tasks take service priority over those arriving later.

IV. Tasks Offloading Scheme

A. Matching Game for Task Assignment

Recently, matching theory (MT) [22] has gained momentum to provide effective solutions to assignment problems. More specifically, the MT establishes mutually beneficial relations between the elements belonging to two opposite sets, taking into consideration the preferences of each element in being assigned to each element of the opposite set. In order to consider the level of satisfaction of each element in being assigned to the element of the opposite set, a list of preferences is created by each element belonging to the two sets. Therefore, the matching game has been formulated between the set of tasks to be offloaded in $\mathcal{U}$ and the set $\mathcal{C}$. Consequently, the metrics of the preferences lists have to be defined, in order to build both the tasks to be offloaded and the network nodes preferences lists. In this reference, the preference lists of each task u in $\mathcal{U}$ on each network node c are created considering the following metric

$$H_u(c) = P_{out,c,u}, \quad \forall c \in \mathcal{C}, \tag{17}$$

where $s_{u,c}$ is the service time experienced by task u on node c. Therefore, on the basis of (17), each $u \in \mathcal{U}$ sorts in ascending order the network nodes $c \in \mathcal{C}$. Differently, the preferences lists of each network nodes c is built considering the deadline associated to each element of $\mathcal{U}$. In fact, each network node c builds its preferences list by sorting in descending order tasks belonging to $\mathcal{U}$, considering their deadlines, i.e.,

$$E_c(u) = t_{d,u}, \forall u \in \mathcal{U}. \tag{18}$$

The matching algorithm, reported in Fig. 2, acts as follows

1) Both the elements in $\mathcal{U}$ and $\mathcal{C}$ create their own preferences lists;
2) Each user $u \in \mathcal{U}$ proposes to be served by the most preferred c_u^*, in accordance with its preference list;
3) Each network node receiving more than one computation requests selects the most preferred c among those proposing. Then, it rejects the other requests;
4) Repeat steps 1)–4) until all the users in $\mathcal{U}$ have been allocated.

Note that the proposed approach does not drop tasks in outage since we assumed as working hypothesis that requests have soft-deadline constraints. It is important to pose the emphasis on the fact that, in our problem, the quantity $\epsilon_{c,i}$ with $c \in \mathcal{U}$, decreases with respect to the original time deadline value as the matching algorithm proceeds, giving rise to preferences lists which dynamically change during the tasks assignment

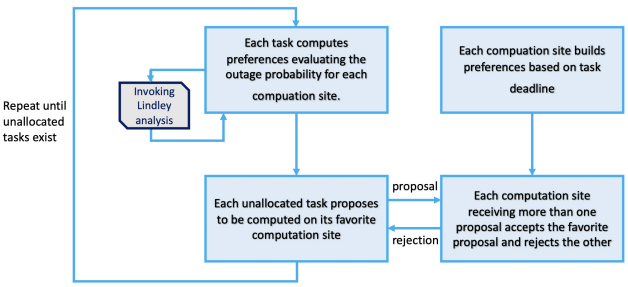

Fig. 2. Diagram of the proposed algorithm, where task preference construction process involves the Lindley's analysis.

[2]The resulting computation service time at node c depends on the computation node capabilities and dimension of u in bytes, however assumed known.

procedure. Such a variation of preferences lists during assignment process implies that the matching game here formulated results to be a matching game with *externalities*, i.e., a game in which the preference lists change in dependence on the allocations performed during the algorithm execution [22]. For this reason, since existing interdependencies and relations among the players' preferences lists exist in the game, preferences lists need to be updated after each algorithm assignment.

Firstly, crucial property of the matching procedure developed requiring investigation is the algorithm termination. With the aim at performing the termination analysis, we focus on a worst case scenario, in which the following assumptions have been considered: i) There is only one network node c; ii) during each iteration, only one task is allocated on the c; In reference to steps 1)–4) of the algorithm previously introduced, and on the basis of working hypothesis i) and ii), the algorithm reaches termination in a number of iteration ι equal to the number of tasks, i.e., in $\iota = |\mathcal{U}|$ steps. Removing hypothesis i) and ii), the corresponding scenario is not worst case, and the algorithm terminates in a number of steps $\iota \leq |\mathcal{U}|$.

B. Complexity Analysis

In order to discuss the complexity of the proposed task offloading strategy, it is important to note that the proposed matching strategy has a computational complexity mainly related to the preference lists creation process of both the parts involved in the game, i.e., the tasks $\mathcal{U}$ and the network nodes $\mathcal{C}$, respectively. Therefore, to build its own preference list, each task orders the elements in the $\mathcal{C}$ set, according to (17), exhibiting a computational complexity, for each task, given by

$$O(|\mathcal{C}| \log |\mathcal{C}|), \tag{19}$$

that considering all the tasks in $\mathcal{U}$ results to be

$$O(|\mathcal{U}||\mathcal{C}| \log |\mathcal{C}|). \tag{20}$$

Similarly, the computational complexity required by $\mathcal{C}$ for sorting elements of $\mathcal{U}$ is

$$O(|\mathcal{C}||\mathcal{U}| \log |\mathcal{U}|). \tag{21}$$

Since in the proposed matching game, the preference lists building processes are the heaviest parts, in terms of temporal complexity. As a consequence, the overall computational complexity is

$$O(|\mathcal{U}||\mathcal{C}| \log |\mathcal{C}|) + O(|\mathcal{C}||\mathcal{U}| \log |\mathcal{U}|). \tag{22}$$

Generally, the number of network nodes is lower than the number of tasks. Then, due to the behavior of the proposed matching algorithm, the network nodes preference lists are built only once, since tasks deadlines do not change over time. Therefore, we have that the overall computational complexity results to be

$$O(|\mathcal{U}||\mathcal{C}| \log |\mathcal{C}|). \tag{23}$$

C. Stability of the Proposed Matching Game

Due to the presence of dependencies among the players' preferences, i.e., externalities, this class of matching games represents a kind of games for which there not exists any matching algorithm able to guarantee convergence to a stable matching [22], [23].

Before stability discussion and analysis, the following *strictly-two-sided exchange-stability* (S2ES) definition is given as a modified version of that originally proposed in [24].

Definition 1. *Let $\mathcal{Z}$ be the final matching produced by the algorithm developed. Let $\mathcal{Z}(u)$ be the network node matched with the task u in the matching $\mathcal{Z}$. The outcome matching $\mathcal{Z}$ is a S2ES matching if there not exists a pair of customers (u_1, u_2) s.t.:*

1) $H_{u_1}(\mathcal{Z}(u_2)) \leq H_{u_1}(\mathcal{Z}(u_1))$ *and*
2) $H_{u_2}(\mathcal{Z}(u_1)) \leq H_{u_2}(\mathcal{Z}(u_2))$ *and*
3) $E_{\mathcal{Z}(u_1)}(u_2) \leq E_{\mathcal{Z}(u_1)}(u_1)$ *and*
4) $E_{\mathcal{Z}(u_2)}(u_1) \leq E_{\mathcal{Z}(u_2)}(u_2)$ *and*
5) $\exists \psi \in \{u_1, u_2\}$ *s.t. at least one of the conditions* $1) - 2)$ *is strictly verified or*
6) $\exists \phi \in \{\mathcal{Z}(u_1), \mathcal{Z}(u_2)\}$ *s.t. at least one of the conditions* $3) - 4)$ *is strictly verified.*

In other words, Definition 1 means that a swap is allowed if it implies an improvement to at least one between the players involved, and all the rest of the elements do not worsen. In order to discuss the stability of the proposed matching algorithm, we suppose the existence of a pair of tasks (u_1, u_2), for which the conditions $1) - 2)$ of Definition 1 results to be satisfied. Furthermore, let u_1 and u_2 be s.t. $\mathcal{Z}(u_1) = c_1$ and $\mathcal{Z}(u_2) = c_2$, respectively. This necessarily means that

$$H_{u_1}(c_2) \leq H_{u_1}(c_1), \tag{24}$$

$$H_{u_2}(c_1) \leq H_{u_2}(c_2). \tag{25}$$

Focusing on condition $5)$ of Definition 1, by (24) and (25), and since the proposed assignment policy does not include any discard strategy, the probability of experiencing a completion time lower than the deadline cannot decrease during the assignment process. Therefore, the preference list of each already matched task cannot change after its assignment. Therefore, we have that at the most $H_{u_1}(c_1) = H_{u_1}(c_2)$ and $H_{u_2}(c_2) = H_{u_2}(c_1)$. As a consequence, condition $5)$ is not verified. In the same way, supposing that the following conditions are true

$$E_{c_1}(u_2) \leq E_{c_1}(u_1), \tag{26}$$

$$E_{c_2}(u_1) \leq E_{c_2}(u_2), \tag{27}$$

we have that $u_2 = u^{\star}_{c_1}$ and $u_1 = u^{\star}_{c_2}$. Since $\mathcal{Z}(u_1) = c_1$ and $\mathcal{Z}(u_2) = c_2$, this means that at the assignment instant conditions $u_1 = u^{\star}_{c_1}$ and $u_2 = u^{\star}_{c_2}$ were true. Due to the fact that once a task is matched its deadline cannot change, (26) and (27) can be verified only if $E_{c_1}(u_2) = E_{c_1}(u_1)$ and $E_{c_2}(u_1) = E_{c_2}(u_2)$. Therefore, the condition $6)$ is not verified and the proposed matching game reaches a configuration satisfying the S2ES property.

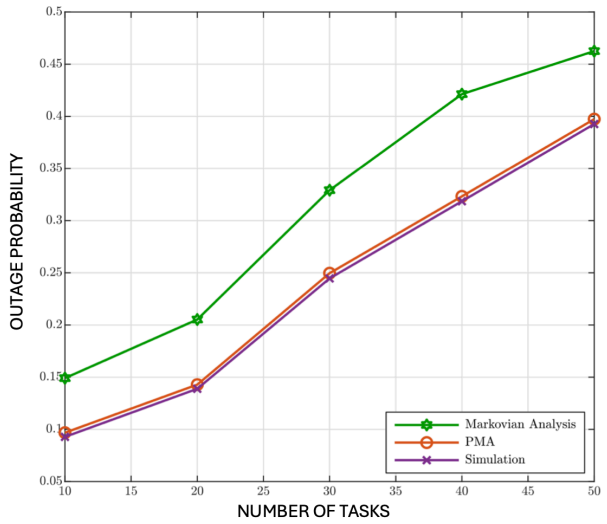

Fig. 3. Outage probability as a function of the number of tasks, in comparison to the Markovian analysis and the experimental simulation curve.

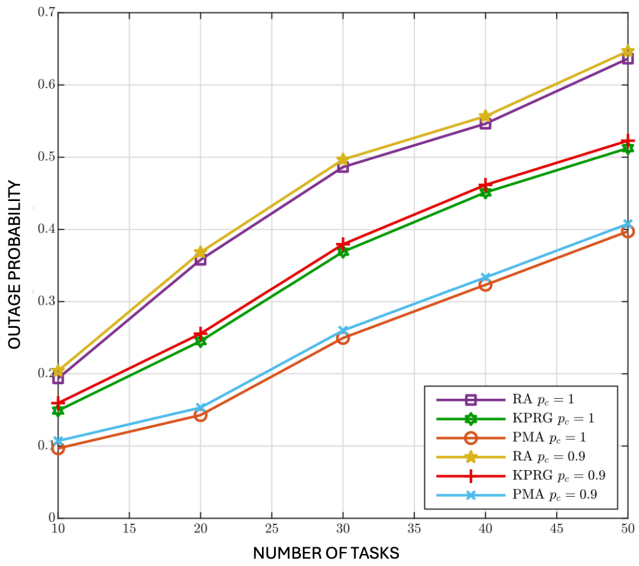

Fig. 4. Outage probability as a function of the number of tasks.

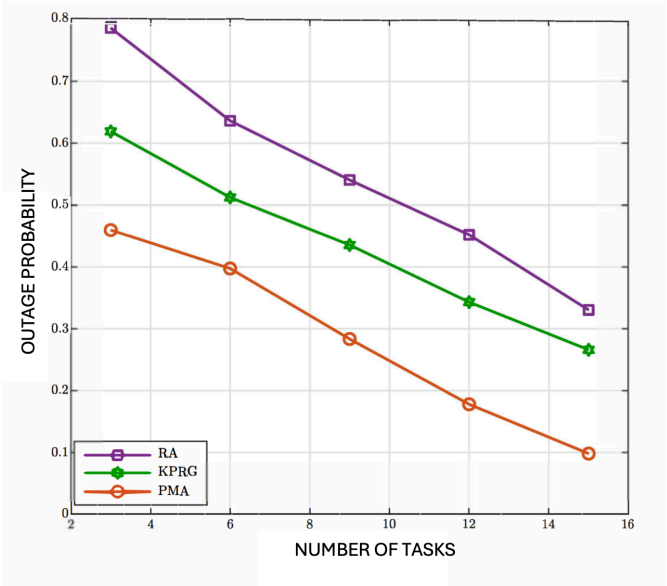

Fig. 5. Outage probability as a function of the ground ENs and 3 UAV-ENs.

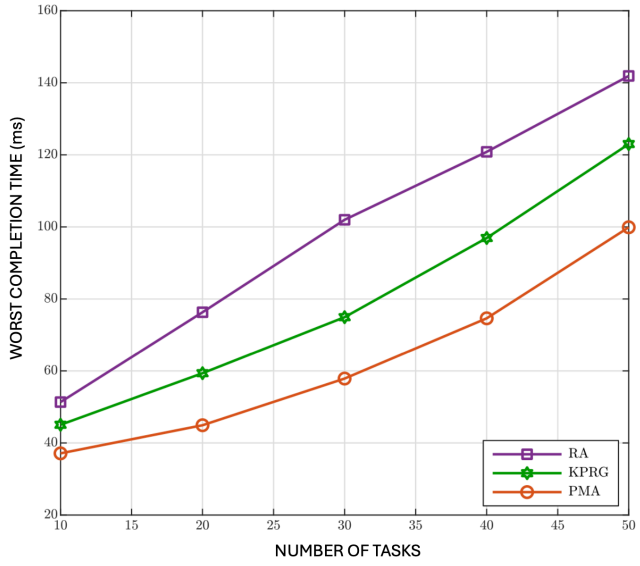

Fig. 6. Worst completion time as a function of the number of tasks.

V. Performance Analysis

This section deals with the performance evaluation of the proposed tasks offloading scheme for the ECI system under the working conditions detailed in Section III-A. We start our analysis by validating the accuracy of the proposed G/G/1 system analysis based on the Lindley's integral in comparison to a classical and more affordable Markov approach based on the assumption of both arrival and service processes as two independent memoryless processes with same mean values as those of the original processes. Furthermore, with the aim at validating the good behavior of the proposed matching algorithm (PMA), we provide performance comparisons with two alternative algorithms: The kolkata paise restaurant game (KPRG) [25], and the random algorithm (RA). In particular, in the RA case, for each task, the network node on which computation is performed is randomly selected with a uniform probability. Conversely, the KPRG game consists of a repeated game in which tasks simultaneously offer to network nodes to be supported for computation, on the basis of the preferences exhibited. Therefore, each task u proposes to the most preferred network EN, i.e., ground EN or UAV-EN, on the basis of its preference list. Then, each network EN, receiving more than one proposal, randomly chooses one of the proposing tasks [22]. As for the system parameters definition, in compliance with [26], we have assumed, without loss of generality, for both arrival and service processes hypoexponential distributions, with pdfs and mean values given in Table I, considering, if not differently specified, 3 ground ENs and 3 UAV-ENs . Moreover, the deadlines

TABLE I
SERVICE AND ARRIVAL CHARACTERIZATION.

$\mathcal{C}$	Service process		Arrival process		Computation speed
	pdf	Mean value	pdf	Mean value	
Ground EN 1	$b(t) = 0.25(e^{-\frac{1}{5}t} - e^{-t})$	1.2	$a(t) = 0.83(e^{-\frac{5}{7}t} - e^{-5t})$	5.7	8 Mbits/s
Ground EN 2	$b(t) = e^{-\frac{1}{7}t} - e^{-\frac{1}{6}t}$	0.30	$a(t) = 0.5(e^{-\frac{1}{7}t} - e^{-\frac{1}{5}t})$	2.24	10 Mbits/s
Ground EN 3	$b(t) = 0.84(e^{-\frac{3}{4}t} - e^{-7t})$	7.75	$a(t) = 0.53(e^{-\frac{1}{2}t} - e^{-8t})$	8.5	5 Mbits/s
UAV-EN 1	$b(t) = 0.52(e^{-\frac{1}{2}t} - e^{-10t})$	10.5	$a(t) = 2.25(e^{-\frac{9}{5}t} - e^{-9t})$	10.8	4 Mbits/s
UAV-EN 2	$b(t) = 1.76(e^{-\frac{3}{2}t} - e^{-10t})$	11.5	$a(t) = 0.51(e^{-20t} - e^{-\frac{1}{2}t})$	20.5	3 Mbits/s
UAV-EN 3	$b(t) = 2.5(e^{-\frac{9}{5}t} - e^{-6t})$	7.8	$a(t) = 1.6(e^{-\frac{7}{5}t} - e^{-11t})$	12.4	2 Mbits/s

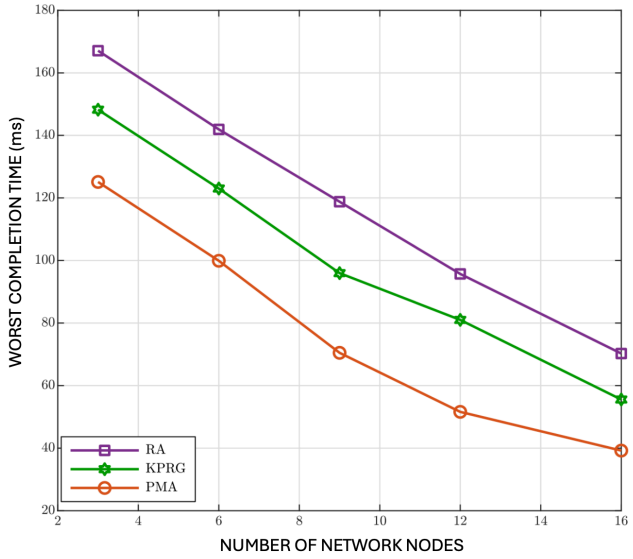

Fig. 7. Worst completion time as a function of the of the ground ENs and 3 UAV-ENs.

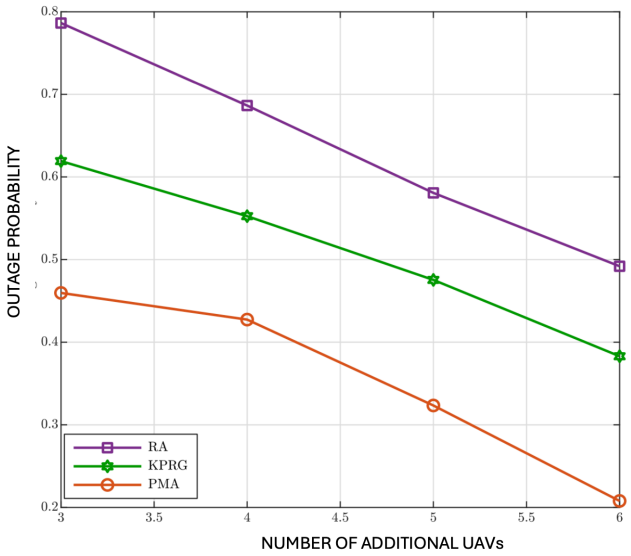

Fig. 8. Outage probability as a function of the number of additional UAVs.

associated to tasks belonging to $\mathcal{U}$ have been assumed heavy tailed distributed within the interval $[30, 110]$ ms. Finally, we have assumed the probability that one UAV can be connected to almost one ground SBS, i.e., p_c, equal to 0.8 and that the network topologies considered in deriving our simulation results are formed by ground ENs and UAV-ENs uniformly selected from the appropriate EN configuration alternatives as reported in Table I. Hence, with the ambition to highlight the advantages of the proposed analytical framework based on the G/G/1 model and Lindley's analysis, we compare in Fig. 3, the obtained analytical predictions in terms of tasks outage probability with simulation results and analytical predictions obtained by resorting to the equivalent M/M/1 model (i.e., where the memoryless arrival an service processes have the same mean values of the actual ones) for which the the waiting time CDF can be derived in a closed form as [4]:

$$W_c^M(t) = 1 - \rho_c e^{-\mu_c(1-\rho_c)t}, \tag{28}$$

where ρ_c is the ratio between the mean arrival and service rates of the equivalent memoryless processes for node c, with $c \in \mathcal{C}$. Finally, in deriving these results we have assumed, as stated before, $p_c = 0.8$. Fig. 3 clearly exhibits the better accuracy of the G/G/1 model based on the Lindley's integral in comparison with the Markov approximation. Moreover, this figure highlights a very good fitting between the obtained analytical predictions and the simulation results that further validates the proposed analysis. As it is evident to note, the outage probability gets worse as the number of tasks increases. This is due to the fact that a higher number of tasks means a greater mean completion time, since the proposed approach does not drop any request. Furthermore, with the aim at confirming the goodness of the proposed PMA task offloading approach in comparison with the RA and KPRG alternatives, (all based on the G/G/1 model), Fig. 4 shows the outage probability as the number of tasks in $\mathcal{U}$ increases and $p_c = 0.8$. As it is evident to note in the figure, the PMA approach exhibits better performance in comparison with the two considered alternatives. Such a trend is due to the fact that the PMA, through the preference lists metrics, minimizes the tasks completion time, that directly impacts the minimization of the outage probability. Differently, both the KPRG and the PMA introduce in the decision-making process a randomness which negatively impacts the system performance. In addition to this, Fig. 5 shows the outage

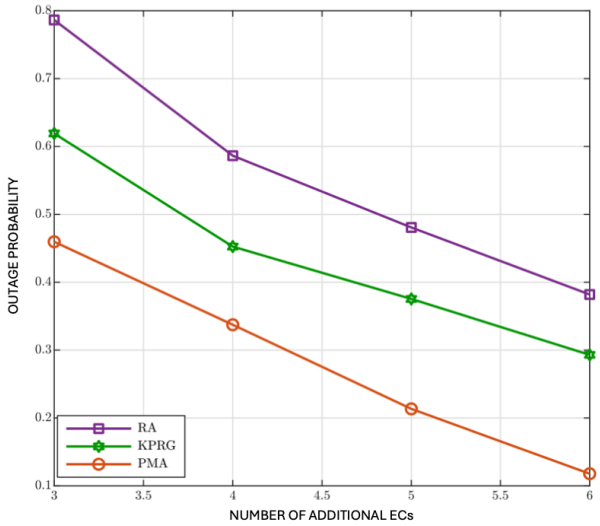

Fig. 9. Outage probability as a function of the number of additional ECs.

probability as the number of ground ENs grows keeping the number of UAV-ENs fixed at 3 for $p_c = 0.8$. Also in this case, the PMA results to be the best strategy in comparison to the alternative schemes taken into account. From Fig. 5, we can note that the outage probability decreases as the number of network nodes increases. This is a consequence of the fact that, as the number of network nodes grows, the computation capability of the system improves, impacting positively on the outage probability behavior. Likewise, Fig. 6 depicts the worst completion time as a function of the number of tasks. Once again, the PMA achieves lower values of the worst completion time, denoting that the proposed matching strategy actually provides a suitable tasks allocation discipline for the problem addressed. The same trend is confirmed in Fig. 7, where the worst completion time is shown as a function of the number of ground ENs, for a number UAV-ENs equal to 3 and $p_c = 0.8$. In addition, Fig. 8 depicts the system performance when we have an additional number of UAVs equal to the value reported in the x-axis. In the same way, Fig. 9 exhibits the performance behavior when an additional number of ECs equal to the value reported in the x-axis is introduced in the network. Consequently, the crossed analysis of Figs. 8 and 9 highlight that to increase the number of ECs lower the outage probability faster than the introduction of the same number of UAVs. This is evident since ECs have a greater computational capability than UAVs.

VI. Conclusions

This paper has focused on a scenario of a hybrid network environment, where computational capabilities are available at the ground ENs and on-board of UAVs. Such network infrastructure has to host computation of a newly arrived set of delay sensitive tasks, assuming the presence of existing nominal flows on the network. The focus of the paper is the minimization of the outage probability of a set of newly arrived tasks under the assumption that the involvement of all the computation nodes in serving nominal data flows are characterized by general independent arrival and service processes. As a consequence, each network node has been modeled as a G/G/1 system and the Lindley's integral analysis, conducted on the basis of the spectrum factorization method [4], has been used in defining the proposed tasks allocation approach based on matching theory principles. Then, the stability of the proposed matching tasks allocation procedure has been also theoretically proved. Finally, with the aim at validating the proposed solution, different allocation procedures have been considered for comparison purposes. The better behavior of the proposed approach has resulted clearly evident from all the results provided here. Future works may include different scheduling policies, also introducing preemption disciplines to prioritize tasks based on the deadline. In addition, another interesting development for this research may be the investigation of per-flow performance bounds involving martingale envelopes theory.

References

[1] P. P. Ray, "A review on 6G for space-air-ground integrated network: Key enablers, open challenges, and future direction," *J. King Saud Univ. - Comput. Information Sciences*, vol. 34, no. 9, pp. 6949–6976, 2022.

[2] Y. Shi and Y. Zhu, "Research on aided reading system of digital library based on text image features and edge computing," *IEEE Access*, vol. 8, pp. 205980–205988, 2020.

[3] V. Kartashevskiy, N. Kireeva, M. Buranova, and L. Chupakhina, "Study of queuing system G/G/1 with an arbitrary distribution of time parameter system," in *Proc. IEEE PIC S&T*, 2015.

[4] L. Kleinrock and K. M. R. Collection, *Queueing Systems, Volume I*, ser. A Wiley-Interscience publication. Wiley, 1974. [Online]. Available: https://books.google.it/books?id=rUbxAAAAMAAJ

[5] H. Huang and A. V. Savkin, "An algorithm of efficient proactive placement of autonomous drones for maximum coverage in cellular networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 994–997, 2018.

[6] B. Shahzaad, A. Bouguettaya, S. Mistry, and A. G. Neiat, "Composing drone-as-a-service (DaaS) for delivery," in *Proc. IEEE ICWS*, 2019.

[7] M. Alwateer, S. W. Loke, and N. Fernando, "Enabling drone services: Drone crowdsourcing and drone scripting," *IEEE Access*, vol. 7, pp. 110035–110049, 2019.

[8] S. Kim and I. Moon, "Traveling salesman problem with a drone station," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 42–52, 2019.

[9] J. Huang, S. Xu, J. Zhang, and Y. Wu, "Resource allocation and 3d deployment of UAVs-assisted MEC network with air-ground cooperation," *Sensors*, vol. 22, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2590

[10] M. Alwateer and S. W. Loke, "A two-layered task servicing model for drone services: Overview and preliminary results," in *Proc. IEEE PerCom Workshops*, 2019.

[11] V. Tarasov, "Transformation of queueing systems into systems with time delay," in *Proc. IEEE PIC S&T*, 2021.

[12] W. He, C. Guo, and X. Wang, "Age of information aware resource allocation and packet sampling control in vehicular networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2245–2249, 2022.

[13] Y. Tang, M. H. Cheung, and T.-M. Lok, "Delay-tolerant UAV-assisted communication: Online trajectory design and user association," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 13137–13151, 2022.

[14] V. Adusumilli and V. TG, "Traffic characterization based stochastic modelling of network-on-chip," *IEEE Trans. Comput.*, vol. 72, no. 4, pp. 1215–1222, 2023.

[15] B. Yang, X. Cao, C. Yuen, and L. Qian, "Offloading optimization in edge computing for deep-learning-enabled target tracking by Internet of UAVs," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9878–9893, 2021.

[16] X. Cao *et al.*, "Reconfigurable intelligent surface-assisted aerial-terrestrial communications via multi-task learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3035–3050, 2021.

[17] F. Jiang *et al.*, "Mars: A DRL-based multi-task resource scheduling framework for UAV with IRS-assisted mobile edge computing system," *IEEE Trans. Cloud Comput.*, vol. 11, no. 4, pp. 3700–3712, 2023.
[18] X. Song, M. Cheng, L. Lei, and Y. Yang, "Multitask and multiobjective joint resource optimization for UAV-assisted air-ground integrated networks under emergency scenarios," *IEEE Internet Things J.*, vol. 10, no. 23, pp. 20342–20357, 2023.
[19] Y. Zhang, J. Guo, Y. Cai, and Y. Wu, "Research on autonomous 6TiSCH network resource demand calculation based on queuing theory," in *Proc. IEEE ICAIBD*, 2022.
[20] F. Chiariotti, B. Soret, and P. Popovski, "Latency and peak age of information in non-preemptive multipath communications," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5336–5352, 2022.
[21] C. Chaccour *et al.*, "Can terahertz provide high-rate reliable low-latency communications for wireless VR?" *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9712–9729, 2022.
[22] D. Manlove, *Algorithmics of matching under preferences*. World Scientific, 2013, vol. 2.
[23] S. Bayat, Y. Li, L. Song, and Z. Han, "Matching theory: Applications in wireless communications," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 103–122, 2016.
[24] E. Bodine-Baron *et al.*, "Peer effects and stability in matching markets," vol. 6982, pp. 117–129, 2011.
[25] T. Park and W. Saad, "Kolkata paise restaurant game for resource allocation in the Internet of things," in *Proc. IEEE ACSSC*, 2017.
[26] K. S. Trivedi, *Probability & Statistics with Reliability, Queuing, and Computer Science Applications*. John Wiley & Sons, Ltd, 07 2016.

Benedetta Picano (M'20) received the B.S. degree in Computer Science, as the M.Sc. degree in Computer Engineering, from the University of Florence, where she received the Ph.D. degree in Information Engineering. She was a Visiting Researcher at the University of Houston. Her research fields include matching theory, nonlinear time series analysis, digital twins, microservices, resource allocation in edge and fog computing infrastructures, and machine learning.

Romano Fantacci (F'05) is a Full Professor of Computer Networks at the University of Florence, Florence, Italy, where he heads the Wireless Networks Research Lab. He received the M.S. degree in Electrical Engineering from the University of Florence, Italy and the Ph.D. degree in Computer Networks from the University of Florence, Italy. His current research interests encompass several fields of wireless engineering and computer communication networking including, in particular, performance evaluation and optimization of wireless networks, emerging generations of wireless standards, cognitive wireless communications and networks, and satellite communications and systems. Dr. Fantacci was elected Fellow of the IEEE in 2005 for contributions to wireless communication networks. He received several awards for his research, including the IEE Benefactor Premium, the 2002 IEEE Distinguished Contributions to Satellite Communications Award, the 2015 IEEE WTC Recognition Award, the IEEE sister society AEIT Young Research Award and the IARIA Best Paper Award, the IEEE IWCMC'16 Best Paper Award and the IEEE Globecom' 16 Best Paper Award. He served as Area Editor for IEEE Trans. Wireless Commun., Associate Editor for IEEE Trans. on Commun., IEEE Trans. Wireless Commun. Regional Editor for IET Communications and Associate Editor for several non-IEEE Technical Journals. He guest edited special issues for IEEE Journals and Magazines and served as Symposium Chair of several IEEE conferences, including VTC, WCNC, PIRMC, ICC and Globecom. Dr. Fantacci currently serves on the Board of Governors of the IEEE sister society AEIT, as Area Editor for IEEE IoT Journal and as member of the Steering Committee of IEEE Wireless Comm. Letters.