A Combined Stochastic Network Calculus and Matching Theory Approach for Computational Offloading in a Heterogenous MEC Environment

Benedetta Picano, *Member, IEEE*, and Romano Fantacci, *Life Fellow, IEEE*

***Abstract*—Nowadays, the functional integration of Unmanned Aerial Vehicles (UAVs) as flying computing nodes with terrestrial networks is rapidly emerging as a promising and viable solution to enhance performance or lower drawbacks arising from unpredictable traffic load congestion occurrences. In particular, this paper considers a UAV-Aided Multiple Access Edge Computing system, in which heterogeneous traffic flows with different quality of service constraints, have to be offloaded on processing nodes consisting of terrestrial and flying edge computing nodes. Towards this goal, the paper proposes a matching algorithm to perform an efficient offloading strategy. In particular, the proposed matching algorithm provides decisions on the basis of per-flow end-to-end delay bounds formulated by resorting to the combined application of stochastic network calculus and martingale envelopes theory. Furthermore, matching stability has been theoretically discussed. Numerical results highlight the validity of the proposed stochastic framework in terms of both reliability, i.e., the probability with which the per-flow end-to-end delay is lower than the corresponding deadline, and its ability to fit the actual network behavior. For comparison purposes, the Boole bound is formulated, and a greedy algorithm is developed to compare the matching strategy designed.**

***Index Terms*—Offloading, matching theory, unmanned aerial vehicle.**

I. Introduction

A. General Considerations

With the advent of the sixth-generation wireless (6G) networks, numerous high-flying challenges have emerged in order to provide reliable, effective, and efficient solutions to handle the expected large volume of heterogeneous tasks computation requests related to novel applications with different quality of service (QoS) constraints and high network responsiveness guarantee, i.e., low delay. In particular, these requirements enable the application of novel technologies in many challenging scenarios, such as smart cities, self-guided vehicles, infrastructure monitoring, and smart grids, to name a few. The huge amount of connectivity, here needed, can be provided by resorting to the upcoming 6G networks by exploiting THz bands [1]. In this way, improvements over network performance can be gained, in perspective, in terms of network responsiveness and end-to-end delay, reliability, coverage, and both spectrum and energy efficiency. Nevertheless, 6G networks are expected to face a vast amount of data traffic and disruptive intensive applications, for which the exclusive usage of terrestrial networks to support Multiple Access Edge computing (MEC) capabilities appears inadequate to properly satisfy the far-reaching traffic demand and the service quality posed by the emerging applications. Therefore, it is through a synergism, carefully orchestrated between terrestrial and non-terrestrial segments of the same network to manage heterogeneous traffic flows, that services delay can be mitigated in a flexible, efficient, and even economic mode [2].

Manuscript received 27 April 2023; revised 31 August 2023 and 16 November 2023; accepted 12 December 2023. Date of publication 14 December 2023; date of current version 15 April 2024. This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE0000001 - program "RESTART"). The associate editor coordinating the review of this article and approving it for publication was T. Inoue. *(Corresponding author: Benedetta Picano.)*

The authors are with the Department of Information Engineering, University of Florence, 50139 Firenze, Italy (e-mail: benedetta.picano@unifi.it; romano.fantacci@unifi.it).

Digital Object Identifier 10.1109/TNSM.2023.3343290

B. Motivations

Nowadays, for terrestrial networks, the emerging MEC paradigm exhibits important advantages due to a distributed computing architecture [3], in which edge nodes (ENs) having computation, storing, and data transmission capabilities are deployed close to the end users [3], at network edges (e.g., in the proximity of 6G small base stations (SBSs)). Due to its distributed nature, MEC aims to reduce the congestion levels and processing times offered by a cloud-based computing architecture. Therefore, MEC represents an effective paradigm to host the computation of task flows stemming from devices needing tasks offloading. In this picture, the functional integration of a UAV, able to host on-board computation, although with reduced capacity in comparison with the terrestrial ENs (T-ENs), allows significant performance improvements. In particular, in such a UAV-Aided MEC system, it is possible to face computation request jams at the ENs, to manage heterogeneous traffic flows, meeting the corresponding variegated QoS constraints [4], [5], [6]. Meanwhile, the design and analysis of proper offloading policies represent a key point in the UAV-Aided MEC system under consideration. To this regard, the time elapsed between the instant in which the device computation request is submitted for processing to the instant of computation request completion, usually referred to as end-to-end (e2e) task flow delay, represents a key metric in a QoS point of view, in particular in the case

of heterogeneous traffic flows. The analysis of e2e delay in UAV-Aided MEC networks has been largely considered since it is recognized as an important methodology to stochastically forecast the network performance [7]. However, the validity of the e2e delay stochastic analysis deeply depends on the class of processes assumed to model arrivals and services. For example, Markovian analysis has been extensively applied, but, sometimes, it is based on too strong and restrictive assumptions to define Markovian approximations of the actual stochastic processes that negatively impact the accuracy of the performed analysis. On the other end, by resorting to a queueing system model with a general characterization for both the arrivals and service processes, i.e., a G/G/1 system model, generally gives rise to a hard computational analysis to be performed in a closed form and, usually limited to the First In First Out (FIFO) queuing policy case [8]. Therefore, most of the previous works on this subject mainly apply the mean e2e delay value analysis, recognized as a simple and widely used approach to study the stochastic processes behavior. Nowadays, the emerging next-generation applications have given rise to traffic flows with quite different statistical characterizations and service requirements that, in turn, have made the mean value queuing analysis no longer applicable. In particular, traditional methods have the well-known key limitation of hiding all of the possible outcomes of randomness, especially considering the true behavior of the system observed regarding specific QoS requirements fulfillment by denying its use for applications with stringent e2e flow computation delay constraints, i.e., Virtual Reality (VR) or even Ultimate VR and run-time monitoring of critical infrastructures. As a consequence, it has become mandatory to identify a more accurate stochastic delay analysis methodology based on the knowledge of the cumulative probability distribution of the parameters of interest instead of their average value. However, the difficulty in developing this type of analysis is well known from the literature, in particular for the case of complex systems such as those of interest here. As a consequence, the ambition of this paper is to propose an affordable analytical approach that provides knowledge of the cumulative probability distribution of the parameters of interest and significantly reduces the analytical complexity of the problem without losing accuracy. Hence, to match this goal and even make it more strict, we have considered as application reliability the probability that the related e2e flow computation delay overcomes a stringent threshold δ. In performing such applications reliability analysis, we have resorted here to the use of the stochastic network calculus (SNC) approach, due to its ability to analyze non-trivial traffic models [9], [10], [11], and to fully catch computer networks behavior in terms of delay[1] [12].

C. Contributions

The main contributions of the paper can be described through the following points

- A performance analysis method for a UAV-Aided MEC system to provide a statistical characterization of the resulting e2e delay for the supported applications, for which we have to consider non-Markov arrivals and service processes. Differently from classical approaches, the provided analytical method can provide stochastic performance bounds with accurate predictions on the actual system behavior;
- The exploitation of the SNC analysis, empowered by the involvement of the martingale envelops [13], [14], to formulate accurate per-flow performance bounds devoted to performing effective offloading decisions. Based on the stochastic worst-case bounds, obtained through the SNC tool integrated with martingale envelops, a matching theory tasks flow offloading algorithm has been developed, providing choices about the flow computation site, i.e., on the T-EN or UAV-EN. To perform such allocation, the per-flow analysis has been conducted by deriving an e2e stochastic bound taking into account the presence of flows already offloaded on the selected computation node and specific computation constraints;
- Numerical simulations to test the performance of the combined SNC-matching theory framework proposed, in terms of adherence to the behavior network bounds formulated with the actual system performance, as well as in terms of validity of the offloading decision-making policy concerning alternative allocation schemes. Finally, to further validate the accuracy of the proposed stochastic bounds, the classical ad wide used Boole bound is considered, for comparison purposes.

The rest of the paper is organized as follows. In Section II an accurate review of the related literature is presented, whereas Section III details the system model and the problem formulation. Section V presents the proposed framework, recalling some key concepts of the SNC for the sake of readability. Performance evaluations are presented in Section VI, while conclusions are drawn in Section VII.

II. Related Works

Recently, UAVs environments have gained significant attention, and many studies have been conducted about them and their interaction with other network landscapes. For example, in [15], authors propose a heuristic to maximize cellular user coverage. In parallel, paper [15] aims at optimizing the drone deployment and communication cost among UAVs. The novel drone-as-a-service market model is advanced in [16] where to satisfy the stringent QoS requirements imposed by users in terms of cost and delay, a service algorithm has been developed and implemented. Then, paper [17] discussed the UAVs limited resources issue. In this picture, the combination and control of multiple UAVs is suggested as a possible solution to overcome these physical limitations. More specifically, programmable crowd-powered drones to create a federated cloud are analyzed, and a scripting language is exploited to properly orchestrate both the flight trajectories of multiple drones and the multi-drone service management. Authors in [18], formulate a mixed integer programming problem, by

[1]Note that the proposed framework considers as reliability metric the probability that the e2e delay is not greater than a given latency constraint δ, the higher it is, the better the performance might be.

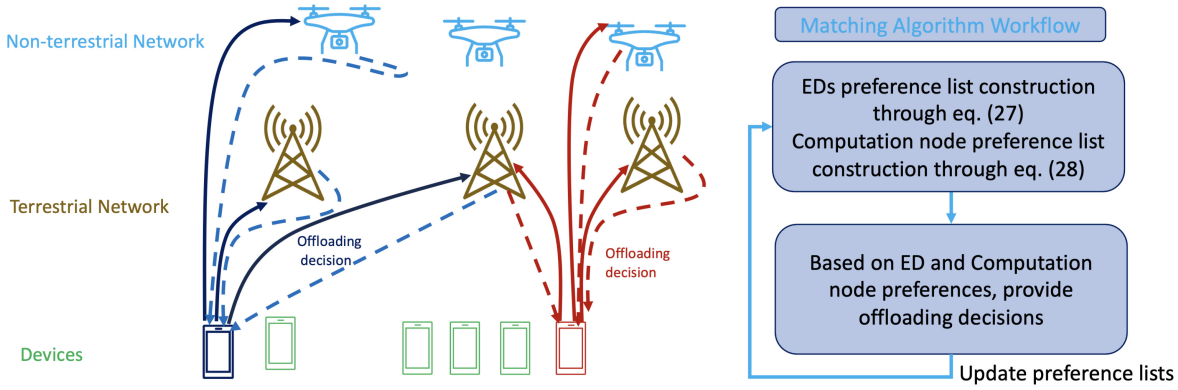

Fig. 1. System scenario.

resorting to a traveling salesman problem with an additional drone station. Consequently, the route distortion problem has been addressed by identifying a lower bound over the number of drones needed to solve the problem object of the analysis. An integrated UAV-assisted MEC network with air–ground cooperation, has been proposed in [19], assuming that both UAV and ground access points have a direct link towards devices and cooperate to process tasks. The main goal is the minimization of the worst delay using the resource allocation procedure, aiming at jointly controlling UAV-device matching, UAV horizontal and vertical position, bandwidth selection, and task splitting. Differently, a two-layered decision-making framework has been developed in [20], promoting the cooperation between one or multiple stations and one or multiple drones, to optimize both profit and travel distance. Interesting machine learning approaches are proposed in [21] and [5], where the offloading problem is addressed by resorting to a deep reinforcement learning approach. In particular, in [21], the authors propose an offloading scheme within a MEC environment to minimize task execution delay and energy consumption, where the UAV is exploited to serve network zones where failures occur. Similarly, in paper [5], both the resource allocation and the offloading problem are addressed, having as objective the long-term minimization of energy and delay in UAV-aided IoT networks. The paper involves clustered multi-UAV and a stochastic game is formulated. Deep reinforcement learning (DRL) is also applied in [22] to solve the surveillance task offloading problem considering a scenario where a UAV swarm is integrated within a mobile edge computing landscape. The paper is devoted to minimizing both the task execution delay and the energy consumption. The DRL is also exploited in [23], where the trajectory of the UAV is optimized along with the computational cost and the network resources. In paper [4] the UAV base station deployment is provided by resorting to artificial intelligence, integrating a novel metric to measure the similarity among user trajectories, to aggregate neural network models with similar costs. Note that all the machine learning-based frameworks are data-driven, having the unavoidable necessity to exploit large amounts of data, often not easily accessible for several reasons (lack/scarcity of datasets supporting experimentation, training, and benchmarking, due to policies of confidentiality, to intrinsic rarity). Differently, the proposed framework is model-driven, proposing an affordable analytical approach that significantly reduces the analytical complexity without losing accuracy. Moreover, due to the ability of the proposed framework to fit the actual behavior of simulations, the proposed approach may be properly extended to generate synthetic data to fed machine learning modules, giving rise to an integrated framework where both model-driven and data-driven approaches act synergistically.

III. Problem Statement

A. System Scenario

As illustrated in Figure 1, we considered a congested T-EN, close to a SBS, hereafter named as tagged SBS, that suffers from the interference of a random set $\mathcal{I}$ of neighboring SBSs. Such as T-EN has to handle an unpredictable overload of service flow computation requests in its service area. To lower the drawbacks arising from this congestion occurrence in a flexible mode (i.e., without requiring a permanent updating of the terrestrial network infrastructure), the functional integration with a UAV having on-board processing capabilities, i.e., acting as a flying EN (UAV-EN) even with some computation and energy limitations, is considered. We have assumed that a set $\mathcal{U} = \{1, \ldots, U\}$ of user devices in the congested area, each one of them requesting a service generating a given flow of tasks. For this reason, the terms device and flow (i.e., service) will be used interchangeably. Each flow has associated a given delay constraint $t_{d,u}$, $u \in \mathcal{U}$. In the congested area, each flow accesses the network through the tagged SBS by means of an individual channel, properly allocated to this purpose. Then, flow computation can be arranged according to the proposed offloading scheme on the most suitable T-EN or UAV-EN sites. We have assumed that the tagged SBS has always available a dedicated reliable channel to communicate with the UAV-EN to offload flow computations, whenever necessary. Once a flow computation is offloaded on the UAV-EN, its computation outcome is sent back to the SBS through a suitable dedicated reliable downlink channel. Finally, regardless of their computation site, i.e., T-EN or UAV-EN, the flow computation outcomes are sent out by the SBS in a broadcast mode to all the devices in the service area through an individual channel. In this way, only those devices that recognize their own ID in the received data packet

TABLE I
LIST OF OUR MAIN NOTATIONS

Notation	Description	Notation	Description
$\mathcal{U}$	set of user devices	$\mathcal{R}_{uav}$	UAV transmission rate
$t_{d,u}$	deadline associated to user u	$\mathbb{P}_{c,u}(W(u) \geq t_{d,u})$	CCPD of the e2e delay
$\mathcal{R}_T$	SBS transmission rate	S^i	service curve
$\mathcal{W}$	bandwidth	A^i	arrival curve
d_M	max distance between user and tagged SBS	d_{UAV}	distance between UAV and tagged SBS
N_0	noise	M_{A^i}	arrival martingale envelope
τ_T	transmission time	M_{S^i}	service martingale envelope
$\mathcal{L}$	packet size	$\mathcal{M}$	supermartingale process
$\hat{k}$	shifted deadline	$\alpha_{u,c}$	offloading decision variable
$V_u(c)$	flow preference list	$E_c(u)$	computation node preference list

are interested in it, others ignore it. In particular, under the assumption of a proper channel allocation, we neglect in the following analysis the influence of the co-channel interference negative effects on the data transmission on dedicated channels from the tagged SBS to the UAV-EN and vice versa. The ideal rate $\mathcal{R}_T$ that assures a reliable data transfer between user devices and the tagged SBS over 6G (i.e., THz) channels, under the condition of an equal power p in transmission, LoS propagation conditions almost surely available, according to [24], results in

$$\mathcal{R}_T = \mathcal{W}\log_2\left(1+\frac{pA_0 d_M^{-2} e^{-K(f)d_M}}{N_0+\sum_{i\in\mathcal{I}} pA_0 d_i^{-2}}\right), \tag{1}$$

where

- $\mathcal{W}$ is the bandwidth of the communication channel;
- d_M is the maximum distance value (worst case) between any user device and the tagged SBS within its service area;
- $\mathcal{I}$ is the random set of interfering signals generated at distance d_i from the receiving end with $i \in \mathcal{I}$;
- $N_0 = \frac{\mathcal{W}\zeta}{4\pi} g_B T_0 + pA_0 d_M^{-2}(1-e^{-K(f)d_M})$ considers both the molecular absorption noise and the Johnson-Nyquist noise at the receiving device site;
- g_B the Boltzmann constant;
- T_0 the temperature in Kelvin;
- ζ the wavelength;
- $K(f)$ the global absorption coefficient of the medium;
- $A_0 = \frac{c^2}{16\pi^2 f^2}$ [25], [26].

The transmission time for a packet formed by $\mathcal{L}$ bits over the ground channels (from user devices to the tagged SBS or vice-versa) results as a random variable having a probability density function defined in [24, Lemma 2] as

$$f(\chi) = \frac{\xi}{\sqrt{2\pi}\sigma_I} e^{-\frac{(I-\mu_I)^2}{2\sigma_I^2}} \tag{2}$$

with

$$\xi = \frac{ln(2)\mathcal{L}pA_0 d_M^{-2} e^{-K(f)d_M} 2^{\mathcal{L}/\mathcal{W}\chi}}{\mathcal{W}\chi^2\left(2^{\mathcal{L}/\mathcal{W}\chi}-1\right)^2}, \tag{3}$$

$$I = \sum_{i\in\mathcal{I}} pA_0 d_i^{-2}, \tag{4}$$

$$\mu_I = p_i A_0 \left(\frac{ln(\delta)-ln(d_M)}{\delta^2-d_M^2}\right)\left(\frac{\pi\delta^2\iota}{2}\right), \tag{5}$$

$$\sigma_I^2 = (pA_0)^2\left(\frac{\pi\delta^2\iota}{2}\right)\left(\frac{1}{2\delta^2 d_M^2}\right), \tag{6}$$

where ι models the intensity of the isotropic homogeneous Matern hardcore point process expresses the SBSs spatial distribution. Hence, under the assumption of task computation request packets formed by a number $\mathcal{L}$ of bits we have that the resulting data packet transmission time is a random variable defined as

$$\tau_T = \mathcal{L}/\mathcal{R}_T \tag{7}$$

Likewise, for the case of a task flow offloaded on the UAV-EN in evaluating the e2e delay we have to take into account that data packet transmissions from the tagged SBS to the UAV-EN and vice-versa are taken place on dedicated channels, for which we have assumed to neglect the negative impact of the co-channel interference. Hence, the data packet transmission rate in both cases results to be a constant, given by:

$$\tau_{uav} = \mathcal{L}/\mathcal{R}_{uav} \tag{8}$$

where we have

$$\mathcal{R}_{uav} = \mathcal{W}\log_2\left(1+\frac{pA_0 d_{UAV}^{-2} e^{-K(f)d_{UAV}}}{N_0}\right) \tag{9}$$

where d_{UAV} is the distance between the tagged SBS and the linked UAV. The e2e delay for a task offloaded

1) On the T-EN: comprises the following contributions:

- the uplink computation request transmission time from user devices to the tagged SBS given by (7);
- the computation system time, i.e., the time spent on board of the T-EN located in the proximity of the SBS (processing time plus waiting time);
- the downlink computation outcome transmission time from the tagged SBS to the destination end user given by (7).

2) On the UAV-EN: we have:

- the uplink offloaded computation request transmission time from user devices to the UAV-EN through the tagged SBS results as the sum of (7) and (8);

- the computation system time, i.e., the time required to perform tasks flows computation on board of the UAV-EN plus the time spent to wait at the UAV-EN site to receive computation;
- the downlink computation outcome transmission time from the UAV-EN to destination user given again by the sum of (7) and (8).

B. Problem Formulation

The main objective of this paper is the maximization of the service reliability provided by the integrated T-EN, UAV-EN system as detailed in what follows. More in-depth, we have to define, throughout the matching game formulated in Section V, an allocation matrix $\mathbf{A} \in \{0,1\}^{U\times(S+V)}$, whose generic element $\alpha_{u,c}$ is equal to 1, if the flow u is offloaded on the network node c, with $c \in \mathcal{C} = \{T\text{-}EN,\ UAV\text{-}EN\}$, or, zero, otherwise. In formal terms, we can define the following optimization problem

$$\min_{\mathbf{A}} \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \sum_{c\in\mathcal{C}} \big[1 - \mathbb{P}_{c,u}\big(W(u) \geq t_{d,u}\big)\alpha_{u,c}\big], \quad (10)$$

where $\mathbb{P}_{c,u}(W(u) \geq t_{d,u})$ represents the probabilistic bound formulated as in (23), and expressing the probability that flow u, offloaded on node c, exceeds the corresponding deadline $t_{d,u}$, i.e., violating the QoS constraint. Specifically, $\mathbb{P}_{c,u}(W(u) \geq t_{d,u})$ expresses the *Complementary Cumulative Distribution* of the e2e delay suffered by packets flow u, i.e., $W(u)$, if it is offloaded on node c. It is important to stress here that the analysis to derive $\mathbb{P}_{c,u}(W(u) \geq t_{d,u})$, is provided in a general form and, hence, it can be applied to any (Markov or non-Markov) service time distribution [27]. In this reference, in Section VI, we have considered the computation node service time as HyperExponentially distributed, according to a widely adopted non-Markov computation time distribution model as discussed in [28].

IV. Stochastic Network Calculus Principles

In what follows, for the sake of readability, we summarize the basic principles of the SNC while the main definitions concerning the martingale envelope will be provided in the Appendix. Note that these principles and the consequent analysis[2] are instrumental to build the preference lists utility functions involved in the matching game outlined in Section V.

Let $T = [a, b]$ be the observation time interval of the system of interest. Intuitively, the cumulative amount of arrivals at the server node during T can be expressed by [7]

$$A_{a,b} = A(a,b) = \sum_{t=a+1}^{b} X_t, \quad (11)$$

where X_t is the number of packets arrived at time t.

From the other hand, $S_{a,b} = S(a,b) = \sum_{t=a+1}^{b} S_t$ describes the corresponding services counted in T.

[2]An in-depth discussion on this issues is available in [27].

According to [27], we can introduce the (*min*, +) convolution operator $\otimes$ and define the bivariate departure process $D(a, b)$ as

$$D(b) = D(0,b) \geq A \otimes S(0,b) := \inf_{0\leq g\leq b}\{A(g) + S(g,b)\}. \quad (12)$$

Then, the delay process $W(b)$ represents the overall time spent in the system by a packet, and which intuitively it is the horizontal distance between the curves $A(n)$ and $D(n)$ [7], [9], [29]. From a theoretical perspective, $W(b)$ is represented by [7], [9], [29]

$$\begin{aligned} W(b) &= W(0,b) \\ &= \inf\{c \geq 0 | A(b-c) \leq D(b)\} \\ &= \inf\left\{c \geq 0 | A(b-c) \leq \inf_{0\leq l\leq n}\{A(l) + S(l,n)\}\right\} \\ &= \inf\left\{c \geq 0 | \sup_{0\leq c\leq n}\{A(c,n) - S(n) \leq 0\}\right\}. \end{aligned} \quad (13)$$

The complementary cumulative distribution function of $W(b)$ results to be [7],

$$\mathbb{P}(W(b) > c) = \mathbb{P}(A(b-c) \geq D(b)). \quad (14)$$

A key point here is that the SNC is a powerful and flexible tool to analyze the e2e delay in computer network systems. Nevertheless, applying the standard SNC envelopes, the resulting bounds are exclusively given on the basis of the arrival processes. Differently, by involving the martingale envelopes, whose definitions are recalled in the Appendix for the convenience of the reader, a proper exponential transformation taking into account both the arrivals and service processes can be formulated.

A. End-to-End Stochastic Bound

In this section, by focusing on the integrated T-EN, UAV-EN system under consideration, we derive the bound on the probability that tasks belonging to a given flow experience an e2e delay greater than the corresponding deadline. Note that this analysis is functional to define the matching offloading scheme in the next Section. We stress here again that the proposed analysis has been conducted assuming that a LoS link is almost surely available to support all the needed communication sessions. However, with the aim at validating our assumption, the obtained analytical predictions will be compared in Section VI with simulation results derived by considering actual propagation conditions, i.e., non-zero probability of having a Non-LoS (NLoS) condition. In performing our analyses, we focus on the tandem systems sketched in Figure 2, which refer to the two possible decisions, i.e., T-EN or UAV-EN, concerning the most suitable computation site for the users' service requests. In particular, the tandem system, illustrated in Figure 2a, is related to the case of tasks computation on the T-EN and it results formed by the following three subsystems: i) the uplink transmission subsystem, involving a given user device and the tagged SBS, i.e., the T-EN, ii) the computation subsystem at the T-EN iii) the downlink transmission subsystem involving the

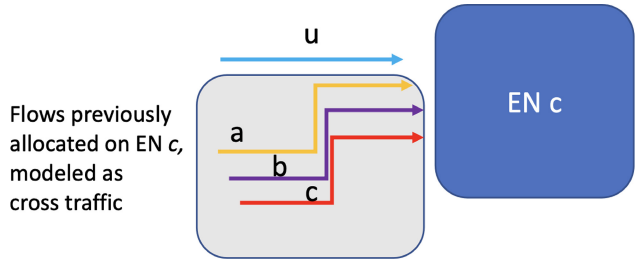

Fig. 2. Flow path offloading models.

T-EN related to the tagged SBS and the interested user device. Similarly, the tandem system related to the case of computation on the UAV-EN is formed by the following subsystems i) uplink transmission subsystem (user device and tagged SBS); ii) air uplink transmission subsystem from the tagged SBS to the UAV; iii) the computation subsystem at the UAV-EN site; iv) downlink air transmission subsystem from the UAV to the tagged SBS; iv) the downlink terrestrial transmission subsystem from the tagged SBS to the requesting user device.

To properly take into account the order with which the tasks flow offloading is performed (see next Section) a static priority (SP) scheduling policy is assumed in performing our analysis. Therefore, computation request packets are served according to the order in which a flow is allocated to the selected computation nodes (T-EN or UAV-EN), i.e., the computation request packets belonging to a flow are allocated first and others are served first. Note that the proposed analysis provides a per-flow, i.e., data tasks sequence sent out by a given device, hereafter named as flow, probabilistic e2e delay bound. In performing our analysis we assume that all the devices have a task, in the form of a packet with the same number of bits independently from the service type, to send out for computation with rate, $\lambda_{\mathcal{H}}$, (packets/s). Hence, by focusing on a given flow, hereafter named as tagged flow, we define as A^1 the cumulative arrivals of the tagged flow and as A^2 the cumulative aggregated arrivals of a generic number U of cross-traffic flows, i.e., $A^2 = \sum_{g=1}^{U-1} C_g$, in which C_g is the g-th traffic flow crossing A^1.

Therefore, as detailed in [30], the SP scheduling for the tagged flow is defined as

$$S^{Type1} = \left[S^{Tot}(m,n) - A^2(m,n)\right]_+ \mathbf{1}_{\hat{n}>x}, \tag{15}$$

in which S^{Tot} is the service curve of the network, obtained applying the min-plus convolution of the service curves of each server, x is a fixed parameter freely chosen, and $\mathbf{1}$ is the indicator function assuming value 1 if the condition $n-m>x$ is satisfied, zero otherwise, accordingly to [7], [9], [12], [30]. Assuming that both the arrivals and services flows admit the martingale envelopes, we will refer in what follows to M_{A^u}, $u \in \mathcal{U}$, for the arrivals processes. Furthermore, we refer to M_{S^i}, $i \in \{2,3\}$, for service processes, where M_{S^2} expresses the martingale service envelope of the computation subsystem, i.e., S^2, and M_{S^3} is referred to the martingale service envelope of the downlink transmission subsystem, i.e., S^3. Consequently, we can conclude that [7], [9], [12][3]

$$\mathbb{P}(W(n) \geq k) \leq \mathbb{P}\left(\sup_{0\leq k\leq n}\left\{A^1(k,n) + A^2(n) - S^1(n) \otimes S^2(n) \otimes S^3(n)\right\} \geq 0\right), \tag{16}$$

$$M_{A^1} \approx h_{A^1}\left(a_n^1\right) e^{\theta\left(A^1(k,n)-(n-k)K^{A^1}\right)}, \tag{17}$$

$$M_{A^2} \approx h_{A^2}\left(a_n^2\right) e^{\theta\left(A^2(k,n)-(n-k)K^{A^2}\right)}, \tag{18}$$

$$M_{S^i} \approx h_{S^i}\left(s_{\tau_i}\right) e^{\theta\left(\tau_i K_s - S^i(\tau_i)\right)}. \tag{19}$$

Therefore, the supermartingale process, follows from the product of (17), (18), and (19) is given by

$$\mathcal{M} = \prod_{j\in\{A^1,A^2,S^1,S^2,S^3\}} M_j. \tag{20}$$

After some algebraic manipulations, and taking into account that

$$\mathbb{E}[\mathcal{M}(k)] = \mathbb{E}\left[M_{A^1}(0)M_{A^2}(0)\prod_{j=1}^{3} M_{S^j}\right], \tag{21}$$

we have

$$\mathbb{E}[\mathcal{M}(k)] \leq \mathbb{E}[M_{A^1}(0)]\mathbb{E}[M_{A^2}(0)]\prod_{j=1}^{3}\mathbb{E}[M_{S^j}(0)]. \tag{22}$$

Then, the martingale bound considering the computation of the tagged flow on the T-EN, results to be

$$\mathbb{P}(W(n) \geq k) \leq e^{-\theta^* kK^{A^1} - kK^{A^2}} B, \tag{23}$$

where k coincides with the deadline associated to the tagged flow, and S^2 represents the service envelope of the service curve for the T-EN, and

$$B = \frac{\mathbb{E}[M_{A^1}(0)]\mathbb{E}[M_{A^2}(0)]\mathbb{E}[M_{S^1}(0)]\mathbb{E}[M_{S^2}(0)]\mathbb{E}[M_{S^3}(0)]}{H}, \tag{24}$$

where $H = \min\{h_{A^1}(a_n^1)h_{S^i}(s_{\tau_i}) : a_n - s_{\tau_i} > 0\}$, and $\theta^* = \sup\{\theta > 0 : K_a \leq K_s\}$, in accordance with [27].

Differently, the bound considering the processing on the UAV-EN, and, hence, taking into account the deterministic contributions due to the uplink and downlink transmission time, τ_{uav}, results to be

$$\mathbb{P}\left(W(n) \geq \hat{k}\right) \leq e^{-\theta^* kK^{A^1} - kK^{A^2}} B, \tag{25}$$

in which $\hat{k} = t_{du} - 2\tau_{uav}$ and S^2 represents the service envelope of the service curve of the UAV-EN, and

$$B = \frac{\mathbb{E}[M_{A^1}(0)]\mathbb{E}[M_{A^2}(0)]\mathbb{E}[M_{S^1}(0)]\mathbb{E}[M_{S^2}(0)]\mathbb{E}[M_{S^3}(0)]}{H}. \tag{26}$$

[3]The complete analysis in reported in [31]

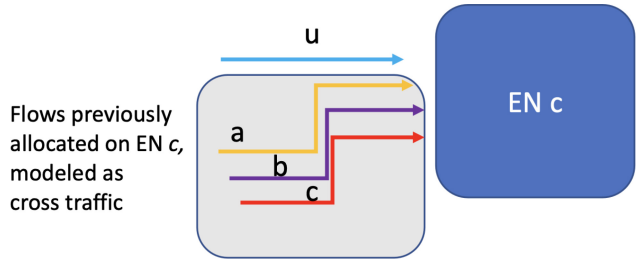

Fig. 3. Flow allocation modeling.

V. Flow Offloading Scheme

This Section illustrates the proposed flows offloading policy based on the Matching Theory. The Matching theory is a well-known mathematical framework able to establish mutually beneficial relations among the elements belonging to two distinct sets. Matching algorithms are particularly useful since, as opposed to more conventional greedy algorithms, the matching games can consider the utility of both the participant sets involved in the game. This characteristic allows us to reach a valuable trade-off between the interests promoted by the two sets. In addition, due to its distributed structure, matching theory, in comparison to standard game theory and Auction theory, based on the full-knowledge about player utilities and actions, does not need full knowledge, since it involves exclusively local utility metrics. Therefore, a matching theory-based algorithm may provide a suitable strategy in offloading environments. As previously anticipated, the matching theory acts based on the preference lists, one for each element belonging to the sets involved in the matching game, aiming at denoting the level of satisfaction of the element in being matched with each element of the opposite set, and vice-versa. In what follows, starting from the preference lists definitions, the matching game is formulated between the two computational nodes alternatives in $\mathcal{C}$ and the set of flows $\mathcal{U}$.

A. Flows Preference List

For each flow u, having deadline $t_{d,u}$, and for each $c \in \mathcal{C}$, the preference list metric $V_u(c)$ of the u-th flow considering c as computation site is given by

$$V_u(c) = \mathbb{P}_{c,u}\big(W(n) \geq t_{d,u}\big), \tag{27}$$

where $\mathbb{P}_{c,u}$ is the probabilistic bound formulated as in (23).

The number of allocated flows, i.e., the flows for which the computation site has been selected, grows as the matching game proceeds. In particular, after the allocation of each flow, i.e., after each algorithm iteration, the preference lists of the unallocated flows have to be updated in order to take into account the impact of the previous allocation decisions. Considering the SNC perspective, as represented also in Figure 3, the presence of previously allocated flows on a node $c \in \mathcal{C}$ is modeled as high-priority traffic flows, in which the priority results to be inversely proportional to the allocation order [32]. Therefore, the flows allocated for the early acquire higher priority than those allocated later during subsequent algorithm steps.

B. Computational Nodes Preference List

Each computational node $c \in \mathcal{C}$, for each $u \in \mathcal{U}$, creates its preference list $E_c(u)$ as

$$E_c(u) = \frac{1}{t_{d,u}}, \tag{28}$$

expressing a higher preference for the flows having a lower deadline.

A modified version of the Gale-Shapley algorithm (GSA) [32], [33] is therefore defined as

1) each flow $u \in \mathcal{U}$ creates the associate preferences list in reference to (27);
2) each computation node $c \in \mathcal{C}$ builds its preference list in accordance with (28);
3) each $c \in \mathcal{C}$ receiving more than one proposal accepts the most favorite one in accordance with its preference list (28). The others are rejected;
4) repeat 1)–3) until all the flow have been allocated on one $c \in \mathcal{C}$.

Since the preferences list of each flow depends on the preferences of other flows, the matching game formulated can be referred to as matching games with *externalities*. In fact, a matching game with externalities is a matching in which there exists interdependence and relations among the players' preferences lists.

C. Stability Analysis

In contrast to standard matching games, the games with externalities are very challenging to handle, since there does not exist any matching algorithm that surely converges into a stable matching.

With the aim of proving the stability of the matching game formulated, the following *strictly-two-sided exchange-stability* (S2ES) definition is introduced, on the basis of the definition previously detailed in [34].

Definition 1: Let $\mathcal{Z}$ be the outcome matching. Let $\mathcal{Z}(i)$ be the c node matched with the u-th flow, in accordance with matching $\mathcal{Z}$. Matching $\mathcal{Z}$ satisfies the S2ES *if* there not exists a pair of flows (u_1, u_2) s.t.:

1) $V_{u_1}(\mathcal{Z}(u_1)) \geq V_{i_1}(\mathcal{Z}(u_2))$ and
2) $V_{u_2}(\mathcal{Z}(u_2)) \geq V_{u_2}(\mathcal{Z}(u_1))$ and
3) $E_{\mathcal{Z}(u_1)}(u_2) \geq E_{\mathcal{Z}(u_1)}(u_1)$ and
4) $E_{\mathcal{Z}(u_2)}(u_1) \geq E_{\mathcal{Z}(u_2)}(u_2)$ and
5) $\exists \psi \in \{u_1, u_2\}$ s.t. at least one of the conditions 1)–2) is strictly verified and
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.

The insight of Definition 1 is that a swap is allowed only if an improvement to at least one between the players involved in the game is achieved, and all the rest of the elements do not get worse. To discuss the stability of the formulated game, we admit the existence of a pair of flows (u_1, u_2), for which

the conditions 1)–2) of Definition 1 are satisfied. Supposing that $\mathcal{Z}(u_1) = c_1$ and $\mathcal{Z}(u_2) = c_2$, we obtain

$$V_{u_1}(c_1) \leq V_{u_1}(c_2), \tag{29}$$
$$V_{u_2}(c_2) \leq V_{u_2}(c_1). \tag{30}$$

In reference to the satisfaction of condition 5) of Definition 1 by (29) and (30).

Since the proposed offloading policy does not include any discard strategy, the delay suffered by the flows allocated cannot change after their assignment, i.e., the delay cannot decrease. Therefore, we have $V_{u_1}(c_1) = V_{u_1}(c_2)$, $V_{u_2}(c_2) = V_{u_2}(c_1)$, and 5) is not satisfied. Vice-versa, considering t_{d,u_1} and t_{d,u_2} the time deadlines corresponding to flows u_1 and u_2, respectively, if c_1 prefers u_2 to u_1, this necessary means that $t_{d,u_2} \leq t_{d,u_1}$. In the same way, if c_2 prefers u_1 instead of u_2, it means that $t_{d,u_1} \leq t_{d,u_2}$. Therefore, we have $t_{d,u_1} = t_{d,u_2}$. In conclusion, neither c_1 nor c_2 obtains benefit in switching, and the condition 6) is not verified, implying that the proposed matching game produces an outcome that satisfies the S2ES property.

D. Complexity Analysis

The complexity of the proposed task offloading strategy is mainly due to the preference list creation by the set $\mathcal{U}$ and the set $\mathcal{C}$. Such a procedure requires invoking the envelope approximation based on the combined SNC-martingale approach, whose complexity is notoriously in the order of $O(|\mathcal{C}|)$ [35]. Consequently, to build its own preference list, each user belonging to $\mathcal{U}$ suffers a complexity given by

$$O(|\mathcal{C}| \log |\mathcal{C}| + O(|\mathcal{C}|)). \tag{31}$$

Extending the analysis to all users in $\mathcal{U}$, the complexity is

$$O(|\mathcal{U}||\mathcal{C}| \log |\mathcal{C}|) + O(|\mathcal{U}||\mathcal{C}|)). \tag{32}$$

Since the computational complexity required by the preference list construction process of elements in $\mathcal{C}$ is exclusively based on (28), the corresponding complexity grows in the order of

$$O(|\mathcal{C}||\mathcal{U}| \log |\mathcal{U}|). \tag{33}$$

To provide a worst-case analysis, we can suppose that the number of algorithm rounds needed to complete allocation is $|\mathcal{U}|$ (i.e., during each round only one user is allocated), concluding that the overall complexity of the proposed framework is

$$O\left(|\mathcal{U}|^2\right). \tag{34}$$

VI. Performance Analysis

This section deals with the performance evaluation of the proposed task offloading approach used in the integrated T-EN, UAV-EN system in terms of the achieved service reliability defined in (10). In addition to this, performance comparisons with a different offloading alternative scheme are provided to validate the good behavior of the proposed solution. In this reference, the network scenario has been set consistent with the parameter values assumed in [24]. As a consequence, simulation results have been derived under actual propagation conditions leading to a NLoS probability of 10^{-3}. Furthermore, we have assumed T-EN and UAV-EN as two heterogeneous computation nodes with an independent identically distributed (*i.i.d.*) computation time assumed hyperexponentially distributed. The mean computation time for UAV-EN is 5.25 ms, whereas that of the T-EN is 3.5 ms. Deadlines have been assumed uniformly distributed in the interval [20, 40] ms. Finally, we have considered a number U of user devices equal to 10, with each of them requesting a specific service. Each user service is related to a flow of task computation requests formed by a data packet with a fixed size equal to 10 Mbits for all the services. For each service flow, a task computation request (i.e., a data packet) is generated with rate $\lambda_{\mathcal{H}}$ (packets/s). In what follows, unless otherwise stated, $\lambda_{\mathcal{H}}$ has been assumed the same for all the devices and equal to 13 packets/s.

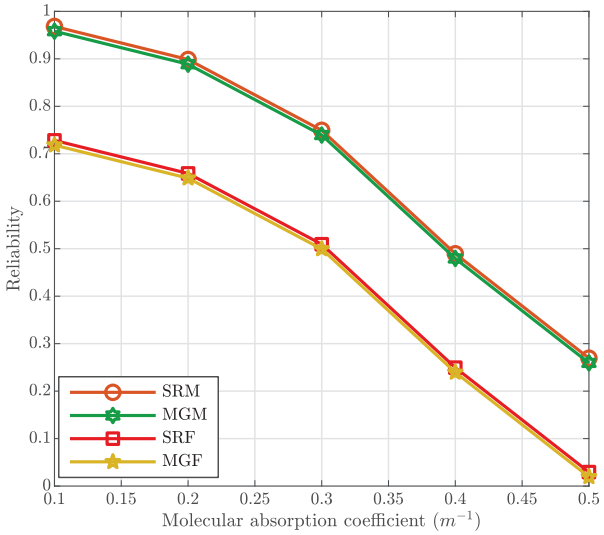

Fig. 4. System Service Reliability as a function of the molecular absorption coefficient.

To validate the proposed analysis and effectiveness of the integrated T-EN, UAV-EN solution, we have compared the analytical predictions (MGM) with results obtained through simulations (SRM). Furthermore, to illustrate the good behavior of the proposed matching game approach, we have implemented an alternative allocation method which operates as follows

- flows are shuffled;
- each flow is assigned to the computation node having the minimum number of flows already allocated;
- the algorithm terminates when all flows are allocated.

The obtained analytical predictions are denoted as MGF while the related simulation results as SGF in the following figures.

Being the investigation of the impact of the use of a 6G network one of the goals of this paper, in Figure 4 we show the influence of the considered THz channel propagation conditions on the system performance in terms of system reliability, i.e., the probability that a given flow experiences an e2e delay lower than its deadline, as a function of the molecular absorption coefficient. Analytical predictions are compared in the figure with simulation results, showing a very good agreement for both the methods considered.

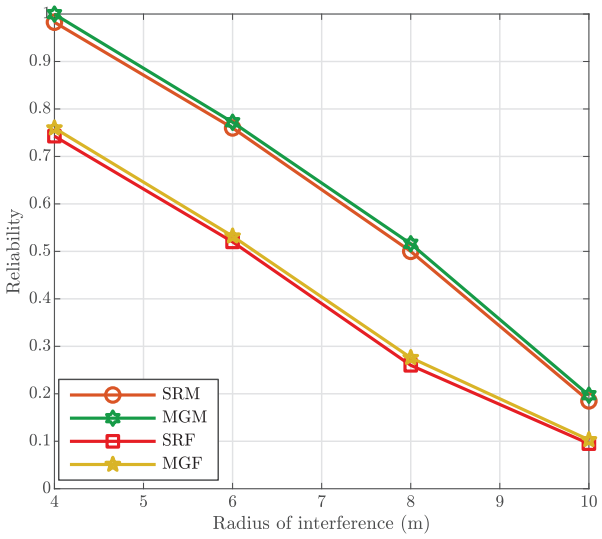

Fig. 5. System Service Reliability as a function of the radius of the interfering region.

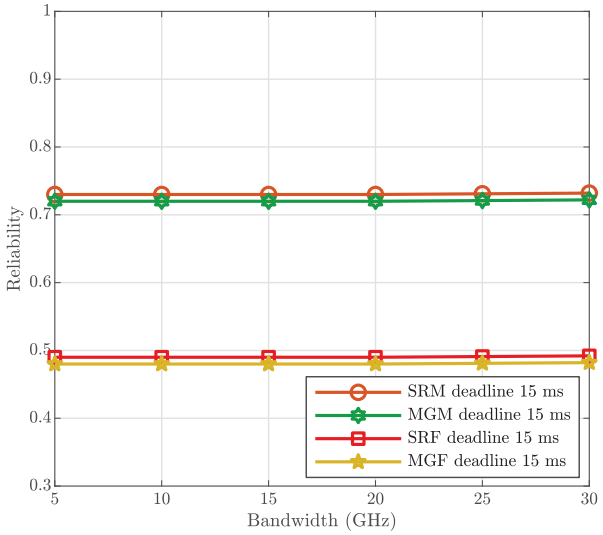

Fig. 7. System Service Reliability as a function of the bandwidth.

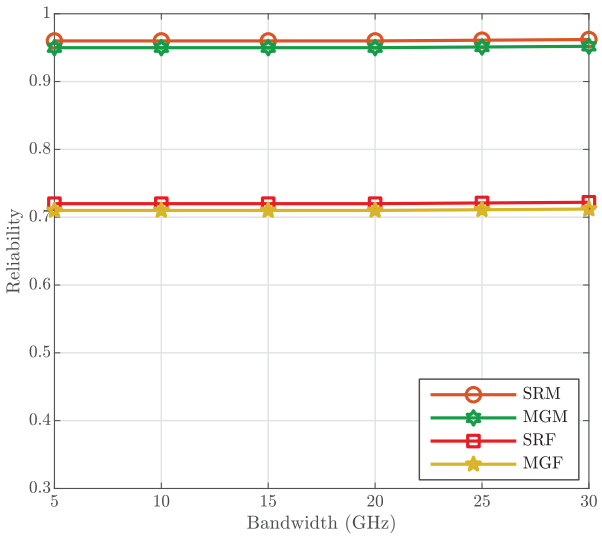

Fig. 6. System Service Reliability as a function of the bandwidth.

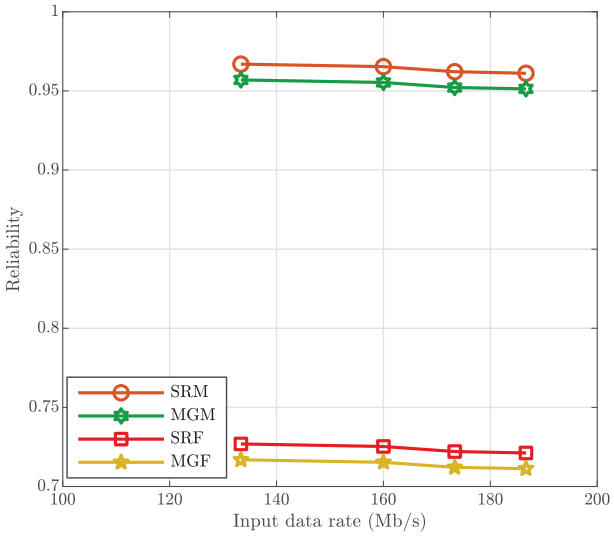

Fig. 8. System Service Reliability as a function of the input data rate.

Moreover, even if a severe impact of the molecular absorption over system performance is evident for both the methods, Figure 4 highlights that the proposed approach exhibits a better behavior. Similarly, Figure 5 depicts the reliability as a function of the radius of the interfering region, i.e., the distance within which SBSs are considered as interfering. Also in this case, even if the achieved performance dramatically degrades as the radius of interference increases for both the considered approaches, a higher resilience to the worst network planning is highlighted for the proposed integrated solution. Furthermore, Figure 6 expresses the behavior of the reliability as a function of the bandwidth of the considered communication channels. Due to the high transmission rates, i.e., low transmission times, supported by the 6G links, the reliability is not significantly impacted by the channel bandwidth, hence, making the computation time at the T-EN or UAV-EN the main critical parameter. Similarly, Figure 7 still exhibits reliability as a function of the bandwidth, considering a mean deadline equal to 15. In comparison to Figure 6, Figure 7 shows a lower level of reliability due to the presence of a more strict deadline. Figure 8 depicts the reliability behavior as a function of $\lambda_{\mathcal{H}}$ given in Mbits/s. From all the previous figures, we have that a good agreement between analytical predictions and simulation results is evident. In particular, this behavior validates the accuracy of the proposed analytical framework and, hence, its suitability to perform an efficient system design and parameters setting without resorting to time-expensive computer simulations. Figure 9 depicts the worst reliability trend (i.e., the reliability of the last flow allocated on the slowest node), as a function of the deadline value. To improve the depth of the analysis provided, we consider the Boole bound [36]. As it is evident to note, results confirm the tightness of the MGM with the SRM curve, compared with the Boole SRM. In the same way, the Boole SRF bound is worse in fitting the actual simulation behavior of the SRF curve. Furthermore, results validate the

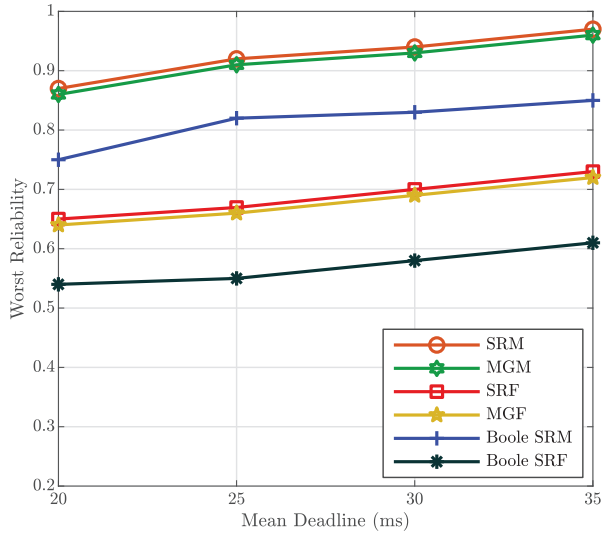

Fig. 9. Worst Service Reliability as a function of the deadline value.

insight according to which a greater deadline value allows to reach higher reliability. For all the values considered, it is also evident that the proposed allocation provides better results, in comparison to the alternative analyzed here. In particular, the achieved performance highlights the ability of the formulated Matching algorithm to properly catch the problem dynamics, exploiting the deadline information. Finally, we conclude this section by highlighting that the clear performance improvements allowed by the considered approach are achieved by resorting to a flexible solution being the UAV dynamically allocated on demand. In particular, the considered integrated solution avoids the need for permanent extensions of the fixed infrastructure, usually no longer necessary in the absence of congestion conditions, and therefore allows a considerable economic advantage (i.e., a reduced network infrastructure deployment cost) without losing the QoS offered to users.

VII. Conclusion

This paper has dealt with the functional integration of a UAV-EN with a T-EN according to the emerging paradigm of a UAV-Aided MEC system to efficiently handle computation load congestion occurrences or to meet strict QoS requirements of the novel applications supported by 6G networks. Towards this goal, a stable matching algorithm has been considered based on the per-flow e2e stochastic bounds analysis formulated by resorting to the SNC and the martingale envelopes [13], [14] to increment the accuracy of the SNC approach. Then, the stability of the proposed matching algorithm has been theoretically proved. Furthermore, performance comparisons with alternative tasks offloading schemes have been also provided to point out the better behavior of the proposed solution. Finally, the good fitting between our analytical predictions with simulation results, derived by considering an actual environment with a non-zero NLoS probability, has validated both the accuracy of the proposed approach and its effectiveness in carrying out the system design and parameters setting without resorting to extensive simulation campaigns.

Due to the ability of the proposed framework to fit the actual behavior of simulations, future works may include the proper extension of the proposed approach to generate synthetic datasets to fed machine learning modules, giving rise to an integrated framework where both model-driven and data-driven approaches act synergistically.

Appendix

In this Appendix the main definitions concerning the martingale envelopes are reported as support to the analytical evaluations provided in the text.

Definition 2 (Submartingale Process): Let $\{\mathcal{F}_n\}_n$ be a filtration such that the stochastic process $\{Y_n\}_n$ is $\mathcal{F}_n$-measurable. $\{Y_n\}_n$ is a submartingale process if, for any time $n \geq 1$, $Y_1, Y_2, \ldots$ satisfy

$$E[|Y_n|] < \infty, \qquad E[Y_{n+1}|\mathcal{F}_n] \geq Y_n. \tag{35}$$

Definition 3 (Martingale Process): The stochastic process $Y_1, Y_2, \ldots$ is a martingale process if, for any time $n \geq 1$, it satisfies

$$E[|Y_n|] < \infty, \qquad E[Y_{n+1}|\mathcal{F}_n] = Y_n. \tag{36}$$

Definition 4 (Supermartingale Process): The stochastic process $Y_1, Y_2, \ldots$ is a supermartingale process if, for any time $n \geq 1$, it satisfies

$$E[|Y_n|] < \infty, \qquad E[Y_{n+1}|\mathcal{F}_n] \leq Y_n. \tag{37}$$

Definition 5 (Arrival Martingale): The arrival process A exhibits martingale arrivals if, for any $\theta > 0, \exists K_a \geq 0$, and $h_a : C(X) \to \mathbb{R}^+$ it satisfies

$$h_a(X_b)e^{\theta(A(b)-bK_a)}, b \geq 1. \tag{38}$$

and the process is supermartingale.

Definition 6 (Service Martingales): The service process S exhibits martingale arrivals if, for any $\theta > 0, \exists K_s \geq 0$, and $h_s : C(X) \to \mathbb{R}^+$ that satisfies

$$h_s(S_b)e^{\theta(bK_2-S(b))}, b \geq 1, \tag{39}$$

and the process is supermartingale.

Definition 7 (Arrivals/Service Martingales): Let R_1, $R_2, \ldots$, be i.i.d random variables, in which the corresponding distributions are nonegative. By assuming generically $A(b) = S(b) = \sum_{g=1}^{b} R_g$, follows that both A and S admit arrival and service martingales, respectively.

References

[1] Y. F. Al-Eryani and E. Hossain, "Delta-OMA (D-OMA): A new method for massive multiple access in 6G," 2019, *arXiv:1901.07100*.
[2] 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. Inf. Sci.*, vol. 34, no. 9, pp. 6949–6976, Oct. 2022.
[3] 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.

[4] Z. Zhao et al., "Predictive UAV base station deployment and service offloading with distributed edge learning," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 3955–3972, Dec. 2021.
[5] A. M. Seid, G. O. Boateng, B. Mareri, G. Sun, and W. Jiang, "Multi-agent DRL for task offloading and resource allocation in multi-UAV enabled IoT edge network," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4531–4547, Dec. 2021.
[6] M. A. Khan et al., "Swarm of UAVs for network management in 6G: A technical review," *IEEE Trans. Netw. Service Manag.*, vol. 20, no. 1, pp. 741–761, Mar. 2023.
[7] T. Liu, J. Li, F. Shut, and Z. Han, "Quality-of-service driven resource allocation based on martingale theory," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.
[8] L. Kleinrock, *Queueing Systems Volume I*. Hoboken, NJ, USA: Wiley, 1974. Accessed: Apr. 2023. [Online]. Available: https://books.google.it/books?id=rUbxAAAAMAAJ
[9] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, 1st Quart., 2015.
[10] Y. Jiang, "A basic stochastic network calculus," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 123–134, Aug. 2006, doi: 10.1145/1151659.1159929.
[11] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1442–1453, Dec. 2007.
[12] M. Fidler, "Survey of deterministic and stochastic service curve models in the network calculus," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 1, pp. 59–86, 1st Quart., 2010.
[13] F. Ciucu and J. Schmitt, "Perspectives on network calculus: No free lunch, but still good value," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 311–322, Aug. 2012, doi: 10.1145/2377677.2377747.
[14] F. Poloczek and F. Ciucu, "Scheduling analysis with martingales," *Perform. Eval.*, vol. 79, pp. 56–72, Sep. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166531614000674
[15] 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, Dec. 2018.
[16] B. Shahzaad, A. Bouguettaya, S. Mistry, and A. G. Neiat, "Composing drone-as-a-service (DaaS) for delivery," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, 2019, pp. 28–32.
[17] M. Alwateer, S. W. Loke, and N. Fernando, "Enabling drone services: Drone crowdsourcing and drone scripting," *IEEE Access*, vol. 7, pp. 110035–110049, 2019.
[18] 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, Jan. 2019.
[19] 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, p. 2590, Mar. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2590
[20] M. Alwateer and S. W. Loke, "A two-layered task servicing model for drone services: Overview and preliminary results," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, 2019, pp. 387–390.
[21] A. M. Seid, G. O. Boateng, S. Anokye, T. Kwantwi, G. Sun, and G. Liu, "Collaborative computation offloading and resource allocation in multi-UAV-assisted IoT networks: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12203–12218, Aug. 2021.
[22] S. M. A. Huda and S. Moh, "Deep reinforcement learning-based computation offloading in UAV swarm-enabled edge computing for surveillance applications," *IEEE Access*, vol. 11, pp. 68269–68285, 2023.
[23] T. Khurshid, W. Ahmed, M. Rehan, R. Ahmad, M. M. Alam, and A. Radwan, "A DRL strategy for optimal resource allocation along with 3D trajectory dynamics in UAV-MEC network," *IEEE Access*, vol. 11, pp. 54664–54678, 2023.
[24] C. Chaccour, M. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless VR?" *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9712–9729, Jun. 2022.
[25] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Risk-based optimization of virtual reality over terahertz reconfigurable intelligent surfaces," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6,
[26] R. Zhang, K. Yang, Q. H. Abbasi, K. A. Qaraqe, and A. Alomainy, "Analytical modelling of the effect of noise on the terahertz in-vivo communication channel for body-centric nano-networks," *Nano Commun. Netw.*, vol. 15, pp. 59–68, Mar. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1878778917300297
[27] F. Poloczek and F. Ciucu, "Service-martingales: Theory and applications to the delay analysis of random access protocols," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2015, pp. 945–953.
[28] K. S. Trivedi, *Probability & Statistics With Reliability, Queuing, and Computer Science Applications*. Hoboken, NJ, USA: Wiley, 2016.
[29] Y. Hu, H. Li, Z. Chang, and Z. Han, "End-to-end backlog and delay bound analysis for multi-hop vehicular Ad Hoc networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6808–6821, Oct. 2017.
[30] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. 14th IEEE Int. Workshop Qual. Service*, 2006, pp. 261–270.
[31] B. Picano and R. Fantacci, "Human-in-the-loop virtual reality offloading scheme in wireless 6G terahertz networks," *Comput. Netw.*, vol. 214, Sep. 2022, Art. no. 109152.
[32] 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, Nov. 2016.
[33] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Mon.*, vol. 69, no. 1, pp. 9–15, 1962. [Online]. Available: http://www.jstor.org/stable/2312726
[34] E. Bodine-Baron, C. Lee, B. Chong, A. Hassibi, and A. Wierman, *Peer Effects and Stability in Matching Markets*. Heidelberg, Germany: Springer, 2011, pp. 117–129.
[35] R. Zippo and G. Stea, "Computationally efficient worst-case analysis of flow-controlled networks with network calculus," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2664–2690, Apr. 2023.
[36] F. Ciucu, "Exponential supermartingales for evaluating end-to-end backlog bounds," *SIGMETRICS Perform. Eval. Rev.*, vol. 35, no. 2, pp. 21–23, Sep. 2007, doi: 10.1145/1330555.1330565.

Benedetta Picano (Member, IEEE) received the B.S. degree in computer science, the M.Sc. degree in computer engineering, and the Ph.D. degree in information engineering from the University of Florence. She was a Visiting Researcher with 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 (Life Fellow, IEEE) received the M.S. degree in electrical engineering and the Ph.D. degree in computer networks from the University of Florence, Italy. He is a Full Professor of Computer Networks with the University of Florence, Florence, Italy, where he heads the Wireless Networks Research Lab. 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. 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 an Area Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the Regional Editor for *IET Communications*, and an associate editor for several non-IEEE Technical Journals. He guest edited special issues for IEEE Journals and Magazines and served as a Symposium Chair for several IEEE conferences, including VTC, WCNC, PIRMC, ICC, and Globecom. He currently serves on the Board of Governors of the IEEE Sister Society AEIT, as an Area Editor for IEEE INTERNET OF THINGS JOURNAL, a member of the Steering Committee of IEEE WIRELESS COMMUNICATIONS LETTERS and the IEEE Comsoc Fellows Evaluation Committee. He was an elected Fellow of the IEEE in 2005 for contributions to wireless communication networks.

Open Access funding provided by 'Università degli Studi di Firenze' within the CRUI CARE Agreement