Seamless Replacement of UAV-BSs Providing Connectivity to the IoT

Hamed Hellaoui[1], Bin Yang[2], Tarik Taleb[3], and Jukka Manner[1]

[1]Aalto University, Communications and Networking Department, Finland. Email: {firstname.lastname}@aalto.fi
[2]Chuzhou University, School of Computer and Information Engineering, China. Email: yangbinchi@gmail.com
[3]University of Oulu, Centre for Wireless Communications, Finland. Email: tarik.taleb@oulu.fi

***Abstract*—This paper considers the scenario of Unmanned Aerial Vehicles (UAVs) acting as flying base stations (UAV-BSs) to provide network connectivity to ground Internet of Things (IoT) devices. More precisely, we investigate the issue where a UAV-BS needs to be replaced by a new one in a seamless way. First, we formulate the issue as an optimization problem aiming to maximize the minimum transmission rate of the served IoT devices during the UAV-BS replacement process. This is translated into jointly optimizing the trajectory of the source UAV-BS (the one to be replaced) and the target UAV-BS (the replacing one), while pushing the IoT devices to seamlessly transfer their connections to the target UAV-BS. We therefore consider a target replacement zone where the UAV-BS replacement can happen, along with IoT connections transfer. Furthermore, we propose a solution based on Deep Reinforcement Learning (DRL). More precisely, we introduce a Multi-Heterogeneous Agent-based approach (MHA-DRL), where two types of agents are considered, namely the UAV-BS agents and the IoT agents. Each agent implements a DQN (Deep Q-Learning) algorithm, where UAV-BS agents learn optimal policies to perform replacement while IoT agents learn optimal policies to transfer their connections to the target UAV-BS. The conducted performance evaluations show that the proposed approach can achieve near optimal optimization.**

***Index Terms*—Unmanned Aerial Vehicles (UAVs), Cellular Networks, Deep Reinforcement Learning (DRL), Multi-Heterogeneous Agent-based DRL (MHA-DRL).**

I. Introduction

Unmanned Aerial Vehicles (UAVs) have been identified as a key enabler of the next generation mobile networks, where UAVs equipped with the adequate radio technology, can act as flying base stations (UAV-BSs) to provide network connectivity to ground devices of the Internet of Things (IoT). This would also allow to support and re-establish network connectivity that has been disrupted due to a natural event of technical failure. UAV-BSs have attracted a lot of attentions from both scientific and industrial communities, which have been translated into different research contributions and implementations/Proof-of-Concepts.

In this paper, we address the issue of replacing UAV-BSs providing network connectivity to ground IoT devices. This event can occur due to different reasons, such as running out of UAV's energy. In this case, it is very crucial to ensure that source UAV-BSs can be replaced by target UAV-BSs seamlessly, so that the services provided by the IoT devices would not be disrupted. Traditionally, the process of changing the serving access network is known as handover (HO) in cellular networks. However, the consideration of UAV-BSs presents new challenges as the access network is moving in the air, contrary to the traditional cellular networks. This underpins other sub-processes, mainly in terms of planning the trajectory of the two concerned UAV-BSs, while transferring the connection of the IoT nodes from a source UAV-BS to a target UAV-BS simultaneously.

The use UAV-BSs to serve ground users has widely been addressed in the literature [1]–[10]. mainly focus on the optimal deployment of the UAVs and optimal resource allocation. This lacks from considering the inevitable scenario of replacing a UAV-BS by another one in a seamless way. To the best of the authors' knowledge, this is the first work to address the problem of seamless replacement of UAV-BSs serving ground IoT devices. In this regard, we first propose a formulation of the problem based on Linear Integer Programming (LIP) to maximize the minimum transmission rate of the served IoT during this process. This is translated into jointly optimizing the trajectory of the source and the target UAV-BS, while pushing the concerned IoT nodes to seamlessly transfer their connections to the target UAV-BS. We also propose a solution based on Deep Reinforcement Learning (DRL). More precisely, we introduce a multi-heterogeneous agent-based approach (MHA-DRL) where two types of agents are considered, namely the UAV-BS agents and the IoT agents. Each agent implements a DQN (Deep Q-Learning) algorithm, where UAV-BS agents learn optimal policies to perform replacement while IoT agents learn optimal policies to transfer their connections to the target UAV-BS.

The rest of the paper is organized as follows. Section II emphasizes with some related works that use UAVs to provide aerial connectivity to ground users. Section III presents the considered system model for the problem of seamless UAV-BSs replacement. This section also introduces a formulation of the problem based on linear optimization. Thereafter, we propose in Section IV a MHA-DRL for the problem of seamless replacement of UAV-BSs. Performance evaluations are provided in Section V. Finally, Section VI concludes this paper.

II. Related Works

UAV-BS has attracted a lot of attentions in the literature. In [1], the authors considered UAV-BSs for wireless data collection from ground IoT devices. Both the deployment of the UAV-BSs and the beamforming design were addressed in this paper. In [2], the authors investigated the joint problem of 3D beamforming design, power allocation, user scheduling and trajectory design for UAV-BS serving ground users. The design of millimeter-wave (mmWave) massive multiple-input multiple-output (MIMO) networks with multiple UAV-BSs has also been investigated in [3]. The authors in [4] elaborated on enhancing the spectral efficiency for UAV-BSs by reducing the communication of command messages used to control the UAV-BSs. The joint uplink-downlink optimization for UAV-BSs is investigated in [5], [6]. To this end, the authors proposed a hybrid-mode multiple access scheme.

The mobility management of a UAV-BS providing connectivity services to a cluster of ground users is investigated in [7]. The paper considered the cases where the geographical characteristics of the cluster and the radio environment are unknown. The joint 3D deployment and power allocation of UAV-BSs for maximizing the system throughput is investigated in [8]. The authors proposed a solution based on DRL to learn the optimal 3D hovering location and power allocation. In another work [9], the authors proposed an optimal placement algorithm for UAV-BSs that maximizes the number of covered users using the minimum transmit power. The energy-efficient 3D placement of a UAV-BS is investigated in [10]. The proposed solution attempts to find the optimal UAV-BS 3D location to support ground users, with minimum UAV-BS energy consumption.

As mentioned previously, many related works on UAV-BSs focus on the issues such as optimal deployment and resource allocation. Unlike existing works, we address the problem of seamless replacement of UAV-BSs providing connectivity to ground IoT nodes. To the best of the authors' knowledge, this is the first work addressing such an issue. The next section introduces the system model and the problem formulation.

III. System Model and Problem Formulation

We consider a UAV-BS scenario to support network connectivity for IoT devices. Let $\mathcal{U}$ and $\mathcal{V}$ denote the set of IoT devices and the set of UAV-BSs, respectively. We also denote by $\mathcal{C}_v$ the set of IoT nodes being served by the UAV-BS $v \in \mathcal{V}$. Thus, we have

$$\underset{v\in\mathcal{V}}{\cup} \mathcal{C}_v = \mathcal{U}, \tag{1}$$

$$\forall v_1, v_2 \in \mathcal{V}; \mathcal{C}_{v_1} \cap \mathcal{C}_{v_2} = \emptyset. \tag{2}$$

We consider replacing one UAV-BS and this process can also be generalized to several UAV-BSs. We use v_s and v_t to denote the source UAV-BS and the target one, respectively. Here, v_s will be replaced by v_t. The general process of replacing UAV-BS is illustrated in Fig. 1. Initially, the source UAV-BS v_s is serving all its IoT nodes ($\mathcal{C}_t = \emptyset$ at the beginning). Thereafter, both the UAV-BSs v_s and v_t start a process of seamless replacement by jointly moving so v_t can take the place of v_s, while transferring the connection of the IoT devices from v_s to v_t ($\mathcal{C}_s = \emptyset$ at the end).

Ensuring seamless replacement of v_s is translated into maintaining enhanced quality-of-service (QoS) for the set of IoT nodes $u \in \mathcal{C}_{v_s}$ during this process. The underlying challenge lies in the fact that UAV-BSs are moving throughout this process. Let us consider the uplink scenario in which data is sent from the IoT nodes to the serving UAV-BSs. Let $p_u = [p_u^n]_{n\in\mathcal{B}_{b(u)}}$ and $h_{uv}(t) = [h_{uv}^n(t)]_{n\in\mathcal{B}_{b(u)}}$ be the transmit power and the channel gain vectors between the IoT node $u \in \mathcal{C}_v$ and its serving UAV-BS $v \in \mathcal{V}$ over the subset of allocated RBs $n \in \mathcal{B}_{b(u)}$. The UAV-BSs use an Orthogonal Frequency Division Multiple Access (OFDMA) technique, which is translated into neglected intra-cell interference. The transmission rate between the IoT node u and its serving UAV-BS v can be expressed as

$$\begin{aligned} r_{uv}(t) &= \sum_{n\in\mathcal{B}_{b(u)}} r_{uv}^n(t) \\ &= \sum_{n\in\mathcal{B}_{b(u)}} W \log_2 \left(1 + \frac{p_u^n h_{uv}^n(t)}{I_{uv}^n(t) + WN_0}\right), \end{aligned} \tag{3}$$

where $I_{uv}^n(t) = \sum_{u'\in\mathcal{U}\setminus\{u\}} p_{u'}^n h_{u'v}^n(t)$ refers to the interference impact from non-served IoT nodes over the same RB, W is the bandwidth of a resource block (RB), while N_0 stands for the noise power. We therefore formulate the problem of maintaining enhanced QoS into maximizing the minimum of the transmission rate of the set of IoT nodes throughout the replacement process.

The trajectory of the UAV-BSs v_s and v_t is planned in the target zone of replacement. We denote by $\mathcal{L}$ the set of possible locations in this zone. In order to plan the trajectory of $v \in \{v_s, v_t\}$, we define the boolean variable $\mathcal{Z}_v^{l,l',t}$ as

$$\mathcal{Z}_v^{l,l',t} = \begin{cases} 1 & \text{if a direct link is formed from the} \\ & \text{location } l \in \mathcal{L} \text{ to the location } \eta(l) \\ & \text{for the UAV-BS } v \in \{v_s, v_t\} \text{ at} \\ & \text{timestamp } t \in \mathcal{T}, \\ & \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

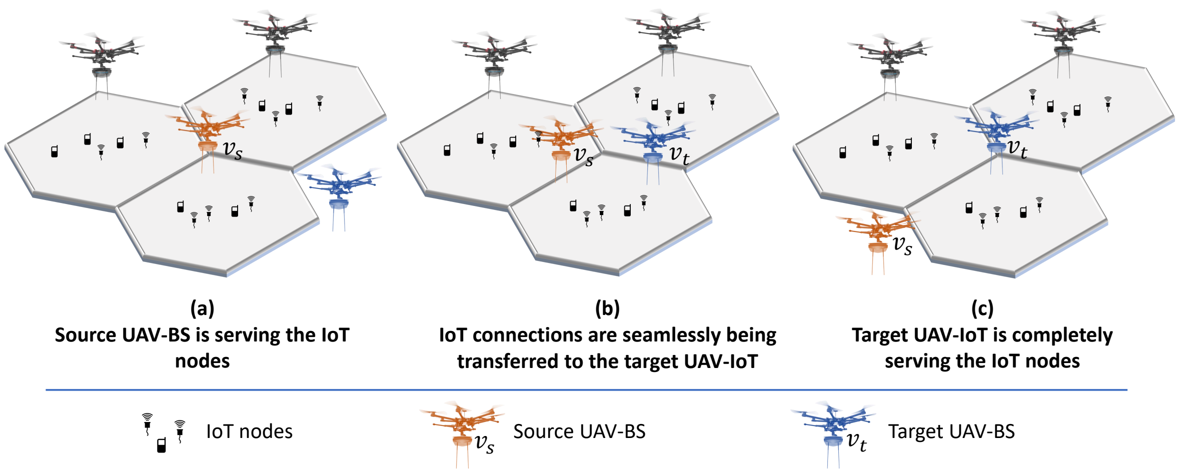
where $\eta(l)$ is the set of neighboring locations of $l \in \mathcal{L}$. We also define $\mathcal{D}(v)$ and $\mathcal{F}(v)$ as the departure and the final locations of the UAV-BS $v \in \{v_s, v_t\}$. We consider that $\mathcal{D}(v_s) = \mathcal{F}(v_t)$ so that v_t would replace v_s. While the UAV-BSs are moving from their initial locations to the target ones, the associated IoT nodes need to seamlessly handover to the target UAV-BS. In order to characterize the operation of transferring the connection of the IoT node $u \in \mathcal{C}_{v_s}$ from v_s to v_t during the replacement process, we define the boolean variable $\mathcal{X}_u^t$ as

$$\mathcal{X}_u^t = \begin{cases} 1 & \text{if the IoT node } u \text{ hands over from } v_s \text{ to} \\ & v_t \text{ at timestamp } t \in \mathcal{T}, \\ & \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Based on (5), we can further characterize the variable $\mathcal{X}_u^t$ as follows:

$$\mathcal{Y}_u^T \triangleq \sum_{t=0}^{T} \mathcal{X}_u^t = \begin{cases} 0 & \text{if } T < \hat{t}, \\ & \\ 1 & \text{if } T \geq \hat{t}, \end{cases} \tag{6}$$

where $\hat{t}$ is the timestamp in which the handover is performed.



(a)
Source UAV-BS is serving the IoT nodes

(b)
IoT connections are seamlessly being transferred to the target UAV-IoT

(c)
Target UAV-IoT is completely serving the IoT nodes

Fig. 1: Illustration of UAV-BS replacement.

The problem of seamless replacement of UAV-BS providing connectivity to the IoT can therefore be formulated as

$$\underset{\{\mathcal{X}_u^t\},\{\mathcal{Z}_v^{l,l',t}\}}{\text{maximize}} \min_{u\in\mathcal{C}_{v_s}} \bar{v}\sum_{l\in\mathcal{L}}\sum_{l'\in\eta(l)}\sum_{t\in\mathcal{T}}\Big(\mathcal{Z}_{v_s}^{l,l',t}(1-\mathcal{Y}_u^t)r_{uv_s}(t) + \mathcal{Z}_{v_t}^{l,l',t}\mathcal{Y}_u^t r_{uv_t}(t)\Big) - \bar{\omega}\sum_{v\in\{v_s,v_t\}}\sum_{l\in\mathcal{L}}\sum_{l'\in\eta(l)}\sum_{t\in\mathcal{T}}\mathcal{Z}_v^{l,l',t}, \tag{7}$$

s.t.

$$\forall v\in\{v_s,v_t\},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\geq 0;\ \ \mathcal{Z}_v^{l,l'}\in\{0,1\}, \tag{8}$$

$$\forall u\in\mathcal{C}_{v_s},\forall t\geq 0;\quad \mathcal{X}_u^t\in\{0,1\}, \tag{9}$$

$$\forall v\in\{v_s,v_t\};\quad \sum_{l'\in\eta(\mathcal{D}(v))}\mathcal{Z}_v^{\mathcal{D}(v),l',0}=1, \tag{10}$$

$$\forall v\in\{v_s,v_t\};\quad \sum_{t\geq 1}\sum_{l\in\eta(\mathcal{F}(v))}\mathcal{Z}_v^{l,\mathcal{F}(v),t}=1, \tag{11}$$

$$\forall v\in\{v_s,v_t\};\quad \sum_{t\geq 1}\sum_{l'\in\eta(\mathcal{F}(v))}\mathcal{Z}_v^{\mathcal{F}(v),l',t}=0, \tag{12}$$

$$\forall v\in\{v_s,v_t\},\forall l\in\mathcal{L};\quad \sum_{t\geq 0}\sum_{l'\in\eta(l)}\mathcal{Z}_v^{l,l',t}\leq 1, \tag{13}$$

$$\forall v\in\{v_s,v_t\},\forall l\in\mathcal{L}\setminus(\mathcal{D}(v)\cup\mathcal{F}(v)),\forall t\in\mathcal{T}; \quad \sum_{l'\in\eta(l)}\mathcal{Z}_v^{l',l,t}=\sum_{l''\in\eta(l)}\mathcal{Z}_v^{l,l'',t+1}, \tag{14}$$

$$\forall l\in\mathcal{L},\forall l',l''\in\eta(l),\forall t\geq 0;\quad \mathcal{Z}_{v_s}^{l',l,t}+\mathcal{Z}_{v_t}^{l'',l,t}\leq 1, \tag{15}$$

$$\forall u\in\mathcal{C}_{v_s};\quad \sum_{t\in\mathcal{T}}\mathcal{X}_u^t=1, \tag{16}$$

$$\forall u\in\mathcal{C}_{v_s},\forall T\geq 1;\quad \mathcal{Y}_u^T=\sum_{t=0}^{T}\mathcal{X}_u^t. \tag{17}$$

The objective of the above optimization problem is to maximize the minimum transmission rate for the IoT nodes, while changing the locations of the two UAV-BSs and transferring the connections from v_s to v_t (the first hand side of (7)). This term includes the variable $\mathcal{Y}_u^t$, which imposes that $\mathcal{Z}_{v_s}^{l,l',t}(1-\mathcal{Y}_u^t)r_{uv_s}(t)$ will be equal to zero after the handover, while $\mathcal{Z}_{v_t}^{l,l',t}\mathcal{Y}_u^t r_{uv_t}(t)$ will be equal to zero before the handover. The above objective function also aims to minimize the paths' length of the two UAV-BSs (the second hand side of (7)). $\bar{v}$ and $\bar{\omega}$ are multi-objective weights used to control the trade-off between reducing the transmission rate the paths' length. Conditions (8) and (9) limit the values of the boolean variables $\mathcal{X}_u^t$ and $\mathcal{Z}_v^{l,l'}$ to the set $\{0,1\}$. Condition (10) ensures that at timestamp 0, the UAV-BSs v_s and v_t will start from their initial locations, $\mathcal{D}(v_s)$ and $\mathcal{D}(v_t)$ respectively. On the other hand, condition (11) guarantees that the UAV-BSs v_s and v_t will reach their target locations, $\mathcal{F}(v_s)$ and $\mathcal{F}(v_t)$ respectively. Condition (12) imposes that the UAV-BSs will stay in their final locations. As for condition (13), it ensures that at most one link can be formed between two locations l and l' at timestamp t for the two UAV-BSs. Furthermore, condition (14) guarantees that the formed path is not interrupted while condition (15) ensures that the two UAV-BSs v_s and v_t will not select the same location at the same time to avoid collision. On the other hand, condition (16) imposes that the IoT node u will handover one and only one time to the UAV-BS v_t. Finally, condition (17) incorporates the definition of $\mathcal{Y}_u^t$.

The above problem formulation is not linear, which is due to the objective function (7) that expresses the product of variables ($\mathcal{Z}_{v_s}^{l,l',t}(1-\mathcal{Y}_u^t)$ and $\mathcal{Z}_{v_t}^{l,l',t}\mathcal{Y}_u^t$). This can be linearized by defining new variables $\mathcal{Q}_u^{l,l',t}$ and $\mathcal{P}_u^{l,l',t}$, which will be imposed to equal $\mathcal{Z}_{v_s}^{l,l',t}(1-\mathcal{Y}_u^t)$ and $\mathcal{Z}_{v_t}^{l,l',t}\mathcal{Y}_u^t$, respectively, by the following set constraints:

$$\begin{cases} \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{Q}_u^{l,l',t}\leq\mathcal{Z}_{v_s}^{l,l',t}, \\ \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{Q}_u^{l,l',t}\leq 1-\mathcal{Y}_u^t, \\ \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{Q}_u^{l,l',t}\geq\mathcal{Z}_{v_s}^{l,l',t}-\mathcal{Y}_u^t, \end{cases} \tag{18}$$

$$\begin{cases} \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{P}_u^{l,l',t}\leq\mathcal{Z}_{v_t}^{l,l',t}, \\ \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{P}_u^{l,l',t}\leq\mathcal{Y}_u^t, \\ \forall u\in\mathcal{C}_{v_s},\forall l\in\mathcal{L},\forall l'\in\eta(l),\forall t\in\mathcal{T}; & \mathcal{P}_u^{l,l',t}\geq\mathcal{Z}_{v_t}^{l,l',t}+\mathcal{Y}_u^t-1. \end{cases} \tag{19}$$

However, the above linear optimization is very complex to be solved by traditional methods (e.g., branch-and-bound), as it involves an important number of variable and constraints. In this paper, we propose a solution based on deep reinforcement learning. Indeed, DRL models can be trained to learn complex tasks. More precisely, we introduce a multi-heterogeneous approach in which both the UAV-BSs and the IoT nodes are considered as agents. Furthermore, we also use the above LIP as a baseline solution to the proposed approach.

IV. A Multi-Heterogeneous Agent-based Deep Reinforcement Learning Approach for Seamless Replacement of UAV-BSs

In this section, we propose a DRL solution for the problem of seamless replacement of UAV-BSs providing network connectivity to ground IoT devices. This solution aims to address the problem formulation introduced in Section III. More precisely, we introduce an approach where heterogeneous agents are considered, namely the UAV-BSs and the IoT nodes. The general architecture of the DRL framework is provided in Fig. 2. At a timestamp t, the different agents (i.e., UAV-BS agents and IoT agents) capture the state of the system s^t which will be used to decide actions (step 1 in Fig. 2). Each agent decides an individual action to be performed, which is reflected by step 2 in Fig. 2 ($a^t_{\mathbf{g}}$ refers to the action performed by the agent $\mathbf{g} \in \mathcal{C}_{v_s} \cup \{v_s, v_t\}$). Thereafter, the agents get the new state of the system, s^t, along with the respective reward (step 3 in Fig. 2). Furthermore, each agent $\mathbf{g}$ also gets in this step the action performed by the other agents, $a^t_{-\mathbf{g}}$, which will be used to learn optimal strategies. Indeed, implementing DRL in a multi-agent environment requires sharing/aggregating the experiences/models between the agents. We also design a replay memory for each agent to store the experiences which will be used to train the neural network model (step 4 in Fig. 2). We detail in what follows the definition of the system state, action space, the system reward and the learning process.

A. System state

The system state is defined in a way to capture the characteristics of the network that allow to achieve the target optimization. At a timestamp t, a system state is defined as

$$s^t = [p_u, h_{uv}(t), \mathcal{D}(\hat{v}), \mathcal{F}(\hat{v}), \mathcal{L}_v(t)]_{u\in\mathcal{U}, \hat{v}\in\{v_s,v_t\}, v\in\mathcal{V}}, \quad (20)$$

where $\mathcal{L}_v(t)$ is the current location of the UAV-BS $v \in \mathcal{V}$.

B. Action space

As mentioned earlier, two types of agents are considered. The action space would therefore differ accordingly. For an IoT node $u \in \mathcal{C}_{v_s}$, an action to be performed at timestamp t is defined as

$$a^t_u \in \{0, 1\}, \quad (21)$$

which indicates whether this node should handover or not to the UAV-BS v_t. As for a UAV-BS $v \in \{v_s, v_t\}$, an action to be performed at timestamp t is defined as

$$a^t_v \in [1, \dots, \eta], \quad (22)$$

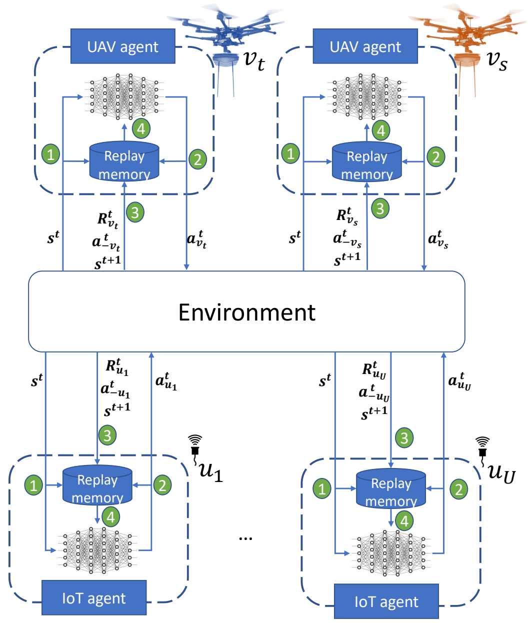


Fig. 2: MHA-DRL framework: two types of heterogeneous agents (i.e., UAV and IoT agents).

where η is the maximum number of neighboring locations. a^t_v therefore allows to select the next location of the UAV-BS v.

C. System reward

The system reward is defined in a way to foster the actions that maximize the objective function. Note that the target optimization is multi-objective as it can be seen in (7). The reward function for an IoT agent $u \in \mathcal{C}_v$ when applying an actions a^t_u and a^t_{-u} (respectively by the agent u and by other agents than u) on a given state s^t is described as

$$\mathcal{R}_u(s^t, a^t_u, a^t_{-u}) = r_{uv}(t). \quad (23)$$

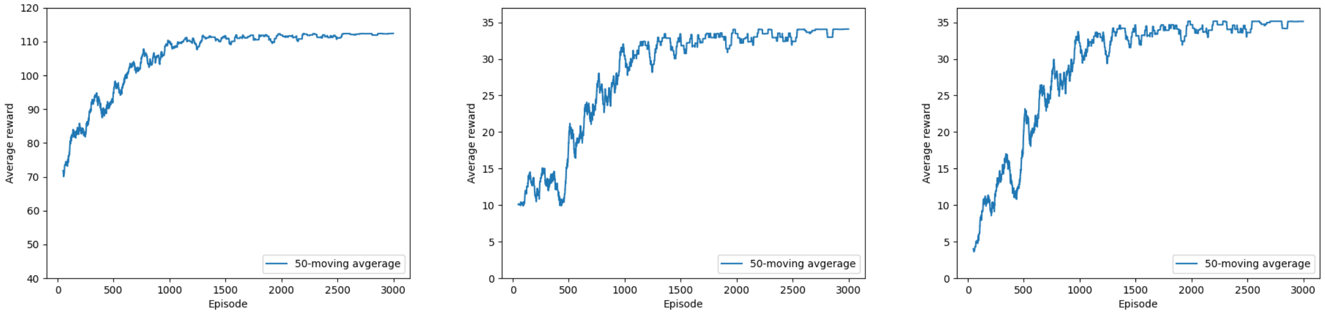
Therefore, increasing the reward of an IoT agent is translated into increasing the associated transmission rate. As for a UAV-BS agent v, the reward function when applying the actions a^t_v and a^t_{-v} (respectively by the the agent v and by other agents than v) is defined as

$$\mathcal{R}_v(s^t, a^t_v, a^t_{-v}) = \begin{cases} \frac{1}{|\mathcal{C}_v|}\sum_{u\in\mathcal{C}_v} r_{uv}(t) & \text{if } v \text{ reaches its destination,} \\ \bar{v}\frac{1}{|\mathcal{C}_v|}\sum_{u\in\mathcal{C}_v} r_{uv}(t) - \bar{\omega} & \text{otherwise.} \end{cases} \quad (24)$$

As we can see in the above equation, increasing the reward for a UAV-BS agent is translated into enhancing the transmission rate of its served IoT nodes while reducing the length of the path.

D. Learning process

In order to learn optimal actions to be performed by the agents, each of the latter implements a DQN algorithm. The underlying objective is to find an optimal policy $\pi_{\mathbf{g}} \in \Pi_{\mathbf{g}}$, for each agent $\mathbf{g} \in \mathcal{C}_{v_s} \cup \{v_s, v_t\}$, that maximizes the long-term



(a) Average reward of the IoT agents (b) Average reward of source UAV-BS agent (c) Average reward of target UAV-BS agent

Fig. 3: Evaluation of the proposed MHA-DRL approach.

reward. This can be expressed using the V-function, $V_{\pi_{\mathbf{g}}}(s)$, as

$$V_{*_{\mathbf{g}}}(s) = \max_{\{\pi_{\mathbf{g}}\}} V_{\pi_{\mathbf{g}}}(s), \tag{25}$$

$$V_{\pi_{\mathbf{g}}}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \tau_{\mathbf{g}}^t \mathcal{R}_{\mathbf{g}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}})|s(0) = s\right]. \tag{26}$$

In the above equation, the symbol $\mathbb{E}[.]$ refers to the expectation operator while $\tau_{\mathbf{g}} \in [0,1]$ reflects a discount factor. Using the Bellman equation, the function $V_{\pi_{\mathbf{g}}}(s)$ can also be written as follows:

$$V_{\pi_{\mathbf{g}}}(s) = \sum_{a_{\mathbf{g}} \in \mathcal{A}_{\mathbf{g}}} \pi_{\mathbf{g}}(a_{\mathbf{g}}|s) \times \underbrace{\left(\mathcal{R}_{\mathbf{g}}(s, a_{\mathbf{g}}, a_{-\mathbf{g}}) + \tau_{\mathbf{g}} \sum_{s' \in \mathcal{S}} P(s'|(s, a_{\mathbf{g}}, a_{-\mathbf{g}})).V_{\pi_{\mathbf{g}}}(s')\right)}_{Q_{\pi_{\mathbf{g}}}(s, a_{\mathbf{g}}, a_{-\mathbf{g}})}, \tag{27}$$

where $\mathcal{A}_{\mathbf{g}}$ is the set of action space for the agent $\mathbf{g}$ (equations (21) and (22)), $a_{\mathbf{g}}$ represents the action taken at the state s by the agent $\mathbf{g}$, $\pi_{\mathbf{g}}(a_{\mathbf{g}}|s)$ refers to the possibility of taking the action $a_{\mathbf{g}}$ when the state is s, while s' denotes the possible resulting states after executing $a_{\mathbf{g}}$ and $a_{-\mathbf{g}}$. The function $Q_{\pi_{\mathbf{g}}}(s, a_{\mathbf{g}}, a_{-\mathbf{g}})$ reflects the Q-function which defines the value of the action $a_{\mathbf{g}}$ taken by the agent $\mathbf{g}$ in the state s under the policy $\pi_{\mathbf{g}}$. The optimal policy can therefore be derived as follows (considering the Bellman optimality equation):

$$V_{*_{\mathbf{g}}}(s^t) = \max_{\{a^t_{\mathbf{g}}\}} Q_*(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}), \tag{28}$$

$$Q_{*_{\mathbf{g}}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}) = \mathcal{R}_{\mathbf{g}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}) + \tau_{\mathbf{g}} \max_{\{a^{t+1}_{\mathbf{g}}, a^{t+1}_{-\mathbf{g}}\}} Q_{*_{\mathbf{g}}}(s^{t+1}, a^{t+1}_{\mathbf{g}} a^{t+1}_{-\mathbf{g}}). \tag{29}$$

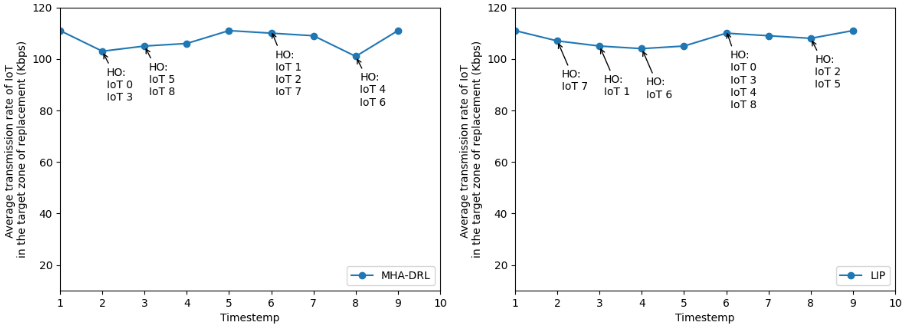
Furthermore, each agent $\mathbf{g}$ implements a deep neural network to estimate the function $Q_{\pi_{\mathbf{g}}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}})$. This can be expressed as $Q_{\pi_{\mathbf{g}}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}) \approx Q_{\pi_{\mathbf{g}}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}, \theta^t_{\mathbf{g}})$, such as $\theta^t_{\mathbf{g}}$ refers to the parameters of the models implemented by the agent $\mathbf{g}$ at timestamp t. The parameters of each model are updated using the gradient decent and the history of experiences stored in the replay memory $\mathcal{M}_{\mathbf{g}}$. This is materialized by iteratively minimizing the loss which is given as

$$\mathcal{L}(\theta^t_{\mathbf{g}}) = \sum_{(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}) \in \mathcal{M}_{\mathbf{g}}} \Big(\mathcal{R}_{\mathbf{g}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}) + \tau_{\mathbf{g}} \max_{\{a^{t+1}_{\mathbf{g}}, a^{t+1}_{-\mathbf{g}}\}} Q_{\pi_{\mathbf{g}}}(s^{t+1}, a^{t+1}_{\mathbf{g}}, a^{t+1}_{-\mathbf{g}}, \theta^{t-1}_{\mathbf{g}}) - Q_{\pi_{\mathbf{g}}}(s^t, a^t_{\mathbf{g}}, a^t_{-\mathbf{g}}, \theta^t_{\mathbf{g}})\Big)^2. \tag{30}$$

V. Performance Evaluations

This section provides the performance evaluation results of the proposed solution. The simulation is performed in a $1000m \times 1000m$ area consisting of 4 cells, where each cell consists of 9 IoT devices served by a UAV-BS. The first cell is the target zone for UAV-BS replacement. We consider a noise power N_0 of $-130dBm$ and a RB bandwidth of $180kHz$. In addition, the discount factor $\tau_{\mathbf{g}}$ is set to 0.9 and the learning rate is set to 0.001.

We first evaluate the performance of the proposed MHA-DRL in terms of learning optimal strategies for performing the UAV-BS replacement process. The obtained results are depicted in Fig. 3. As we can see, the different agents are able to learn strategies allowing to maximize the reward values. For the IoT agents (Fig. 3 (a)), increasing the reward value is translated into maximizing the transmission rate of the corresponding IoTs throughout the UAV-BS replacement process. Indeed, the reward function for an IoT agent is expressed based on the transmission rate of the IoT nodes (See equation (23)). Consequently, the IoT agents learn the optimal time to perform handover. As for the source and the target UAV-BS agents (Fig. 3 (b) and Fig. 3 (c), respectively), increasing the reward values implies reducing the distance to the target location, while ensuring that the selected trajectory maintains a increased transmission rate for the IoT devices connected to this UAV-BS (equation 24 provides the reward function for a UAV-BS agent). We can also see from this evaluation that the source and the target UAV-BSs reach less reward value compared to the IoT agents. This is due to the fact that reward value for the UAV-BS agents decreases with the distance taken to reach the target destination (so to avoid long paths).



(a) MHA-DRL

(b) LIP

Fig. 4: Evaluation of the UAV-BS replacement process for the proposed MHA-DRL against LIP (optimal solution).

Furthermore, we compare the proposed MHA-DRL approach with a baseline solution (optimal solution). The latter is reflected in the LIP solution proposed in Section III. To this end, we use Gurobi [11] as a solver. The obtained results are depicted in Fig. 4. In terms of paths' length, the two solutions achieve the same number of waypoints (9 steps performed throughout 9 timestamps). Furthermore, the two solutions maintain an average transmission rate, for the set of IoT in the target zone of replacement, which is similar to the initial value ($111Kbps$). This shows that the execution of the UAV-replacement process has been implemented in a seamless way and proves that the proposed MHA-DRL achieves near optimal optimization. Fig. 4 also shows the timestamps in which the handover operation has been executed for each of the 9 IoT devices in the target zone (dubbed IoT 0, $\ldots$, IoT 8). We can see that the handover is executed at different timestamps for each IoT node, considering the two solutions. Indeed, the decision for handover is made to maintain enhanced transmission rate for the corresponding IoT node and is not linked to a specific timestamp. On the other hand, the execution of the above LIP in the Gurobi solver is translated into 92955 constraints and 32361 variables. This makes the proposed MHA-DRL much faster and practical compared to the LIP optimization. More precisely, the solving the above LIP in a x86_64 machine with 8 CPUs of $2397.224MHz$ requires $3374s$, while the execution of one episode of the HMA-DRL approach in the same machine requires in average $0.7ms$.

VI. Conclusion

This paper investigated the issue of replacing UAV-BSs providing connectivity to the IoT. We formulated it as an optimization problem for maximizing the minimum transmission rate of the served IoT devices by jointly optimizing the trajectory of the UAV-BSs (source and target), while performing IoT handover. To solve the challenging optimization problem, we proposed a solution based on DRL, with which we adopted a Multi-Heterogeneous Agent approach including two types of agents, i.e., the IoT agents aiming to select optimal handover time, and UAV-BS agents aiming to select optimal paths. The performance evaluations show that the agents are able to learn optimal strategies allowing to execute the process of UAV-BS replacement in a seamless way. Meanwhile, they also show that the proposed MHA-DRL achieves similar results compared to the optimal solution, while being able to be executed in a short time.

Acknowledgment

This work was supported by the European Union's Horizon 2020 Research and Innovation Program through the 5G!Drones Project under Grant No. 857031.

References

[1] Z. Xiao, H. Dong, L. Bai, D. O. Wu, and X.-G. Xia, "Unmanned Aerial Vehicle Base Station (UAV-BS) Deployment With Millimeter-Wave Beamforming," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1336–1349, 2020.

[2] K. Li, X. Zhu, Y. Jiang, and F.-C. Zheng, "Closed-Form Beamforming Aided Joint Optimization for Spectrum- and Energy-Efficient UAV-BS Networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[3] L. Zhu, J. Zhang, Z. Xiao, and R. Schober, "Optimization of Multi-UAV-BS Aided Millimeter-Wave Massive MIMO Networks," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[4] H. Hellaoui, O. Bekkouche, M. Bagaa, and T. Taleb, "Aerial Control System for Spectrum Efficiency in UAV-to-Cellular Communications," *IEEE Communications Magazine*, vol. 56, no. 10, pp. 108–113, 2018.

[5] H. Zeng, X. Zhu, Y. Jiang, Z. Wei, and Y. Hao, "Hybrid-Mode Multiple Access for UAV-BS Assisted Communications with UL-DL Rate Balancing," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[6] H. Zeng, X. Zhu, Y. Jiang, Z. Wei, S. Sun, and X. Xiong, "Toward UL-DL Rate Balancing: Joint Resource Allocation and Hybrid-Mode Multiple Access for UAV-BS Assisted Communication Systems," *IEEE Transactions on Communications*, pp. 1–1, 2022.

[7] M. G. Khoshkholgh and H. Yanikomeroglu, "RSS-Based UAV-BS 3-D Mobility Management via Policy Gradient Deep Reinforcement Learning," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

[8] M. Zhang, S. Fu, and Q. Fan, "Joint 3D Deployment and Power Allocation for UAV-BS: A Deep Reinforcement Learning Approach," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2309–2312, 2021.

[9] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D Placement of an Unmanned Aerial Vehicle Base Station (UAV-BS) for Energy-Efficient Maximal Coverage," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 434–437, 2017.

[10] J. You, S. Jung, J. Seo, and J. Kang, "Energy-Efficient 3-D Placement of an Unmanned Aerial Vehicle Base Station With Antenna Tilting," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1323–1327, 2020.

[11] Gurobi, "Gurobi optimization," http://www.gurobi.com/, [Online].