Harnessing UAVs for Fair 5G Bandwidth Allocation in Vehicular Communication via Deep Reinforcement Learning

Tingting Yuan, Christian Esteve Rothenberg, Katia Obraczka, Chadi Barakat, Thierry Turletti

Harnessing UAVs for Fair 5G Bandwidth Allocation in Vehicular Communication via Deep Reinforcement Learning

Tingting Yuan*, Christian Esteve Rothenberg†, Katia Obraczka‡, Chadi Barakat*, Thierry Turletti*
*Inria, Université Côte d'Azur, France
†University of Campinas, Brazil
‡UC Santa Cruz, CA, USA

***Abstract*—Terrestrial wireless infrastructure-based networks do not always guarantee that their resources will be shared uniformly by nodes in vehicular networks mostly due to the uneven and dynamic distribution of vehicles in the network as well as path loss effects. In this paper, we leverage multiple fifth-generation (5G) unmanned aerial vehicles (UAVs) to enhance network resource allocation among vehicles by positioning UAVs on-demand as "flying communication infrastructure". We propose a deep reinforcement learning (DRL) approach to determine the position of UAVs in order to improve the fairness and efficiency of network resource allocation while considering the UAVs' flying range, communication range, and limited energy resources. We use a parametric fairness function for resource allocation that can be tuned to reach different allocation objectives ranging from maximizing the total throughput of vehicles, maximizing minimum throughput, as well as achieving proportional bandwidth allocation. Simulation results show that the proposed DRL approach to UAV positioning can help improve network resource allocation according to the targeted fairness objective.**

***Index Terms*—Unmanned aerial vehicles (UAV), fifth-generation (5G), fairness, deep reinforcement learning (DRL).**

I. Introduction

Even though it is expected that wireless communication will become increasingly more ubiquitous and accessible in urban regions through base stations (BSs), road-side units (RSUs), the quality of service experienced by wireless users may vary significantly especially due to the uneven and highly dynamic distribution of vehicles as well as communication channel impairments (e.g., path loss). Employing unmanned aerial vehicles (UAVs) to complement existing wireless communication infrastructure deployments, in particular, to assist in vehicular communication, has been the topic of a number of recent research efforts [1]–[3]. Their ability to provide line-of-sight (LoS) links, coupled with their agility and rapid deployability, enable UAVs to be deployed on-demand to serve as "flying access points". As such, they can assist the installed communication infrastructure in providing adequate communication services, especially in highly dynamic scenarios such as vehicular networks in urban, suburban, and rural regions. This holds particularly for fifth-generation (5G) and beyond networks, in which LoS communication is critical due to diffraction and material penetration effects which incur greater attenuation [4].

However, deploying UAVs to provide sufficient communication coverage to vehicles "anywhere, anytime" is quite challenging given that: UAVs have limited communication range, are power-constrained [5], and there will only a limited number of UAVs to cover a certain region. Additionally, UAV location is in three-dimensional (3D) space, which means that their communication channel characteristics vary not only with their location relative to the ground but also their altitude. Therefore, UAV placement over time needs to consider a number of aspects including vehicle location and their bandwidth requirements, the capacity of the local terrestrial communication infrastructure, as well as the UAV energy budget and communication capability.

Some previous efforts on UAV placement try to maximize aggregate throughput [6]–[8], while some other proposals aim to maximize the number of users under service [9]. However, these approaches are either specific in their optimization target or may create scenarios of unfair resource allocation as they tend to provide resources to well-connected vehicles while ignoring or under-serving poorly connected ones.

Recently, computational intelligence, in particular machine learning (ML), has gained significant traction as a powerful tool to address challenges posed by new, more complex problems in a variety of disciplines and domains. The widespread use of ML techniques has been fueled in part by the ever-increasing availability of computing power. As networks and their applications become increasingly more complex and heterogeneous, they too can benefit from ML approaches which can learn and adapt to network- and application dynamics automatically and autonomously. Some ML techniques do not require a priori knowledge of the operating environment, but they acquire this knowledge as they operate and adjust accordingly without the need for complex mathematical models of the system. In particular, deep reinforcement learning (DRL) [10], [11] has been viewed as a promising solution for dynamic UAV placement [8], [12] since it is well-suited to tackle problems that require longer-term planning using high-dimensional observations; additionally, it uses powerful deep neural networks to build a higher-level understanding of the target environment with limited- or even no prior knowledge.

In this paper, we propose a DRL-based approach to perform optimal, real-time UAV placement in order to achieve fair and

efficient bandwidth allocation in 5G vehicular networks. To this end, we model the problem of dynamic UAV placement as an optimization problem whose goal is to maximize a global objective fairness function calculated over the communication bandwidth allocated to vehicles while accounting for *(i)* the aggregate communication resources, i.e., bandwidth, provided by ground wireless communication infrastructure and deployed UAVs, *(ii)* the time-varying vehicles' location, and *(iii)* the UAVs' energy budget. By employing a parametric fairness function, which is inspired by previous work on Internet resource allocation (e.g., [13]), as an objective function, UAVs are dynamically placed in such a way as to realize different allocation strategies including maximizing aggregate throughput of vehicles, maximizing minimum vehicle throughput, as well as achieving proportional fairness over vehicle's throughput. We evaluate the performance of the proposed DRL-based UAV placement framework by means of simulations driven by real-world vehicle mobility traces.

The remainder of the paper is organized as follows. Section II provides an overview of related work. Section III describes our system model including the underlying communication channel and deployed infrastructure, UAV energy consumption, as well as vehicle spatial and temporal distribution. Section IV formulates UAV placement as a non-linear optimization problem and uses time discretization to approximate. Section V introduces the proposed DRL-based method to solve the proposed problem. Section VI describes our experimental methodology and presents results. Finally, Section VII concludes the paper.

II. Related Work

Optimizing dynamic placement of UAVs to improve the performance of communication networks is complex and challenging and has been the focus of a number of research efforts in recent years. Early work focused on single UAV placement [7], [9]. Other efforts explored the use of UAV swarms. For example, [14] proposes to minimize the number of UAVs to ensure coverage of all users, and [15] suggests to maximize the total coverage areas. The placement problem is even more complex when multiple UAVs are considered, with new challenges on UAVs' cooperation and placement algorithm efficiency.

Considering both high-dimensional state-space and time-varying environments, ML-based technologies have been recently utilized for solving challenging problems in UAV placement [16], [17]. DRL provides a promising solution as it can handle efficiently a high-dimensional state-space and time-varying environment. Recently, a few works have been conducted to investigate the use of DRL for the placement of multiple UAVs. In [6], a DRL-based approach for UAV placement that complements terrestrial communication to maximize total network throughput is proposed, and in [8], DRL is also used to place UAVs to maximize aggregate network throughput while simultaneously balancing traffic load across UAVs. The work reported in [18] explores DRL to maintain acceptable QoS for each vehicle and minimize the number of UAVs. The authors of [12] study energy-efficient UAV placement with DRL considering the fairness on coverage time. This work tries to even the cells' serving time by UAVs, without considering the users' communication quality.

Unlike previous work, we consider a novel and important objective for dynamic UAV placement, namely network resource allocation fairness such that vehicles have access to a fair share of the network bandwidth. To achieve this goal, we use a parametric fairness function inspired by early work on Internet resource allocation [13] as objective and maximize it over the offered load from vehicles.

III. System Model

As illustrated in Fig. 1, we consider a scenario of a limited region and with some terrestrial communication infrastructures deployed. UAVs, indexed by $U = \{u_1, \ldots, u_N\}$, are air-wireless access points serving vehicles on the ground. Thus, the communication infrastructures of this scenario includes a set of BS, RSUs and UAVs, which we denoted as $I = \{S, R, U\}$. The distance between two points in the space is defined as $d_{i,j} = \|l_i - l_j\|$, with $\|\cdot\|$ denoting the Euclidean norm, and l_i is the location of one device. Besides, there are charging stations with location l^{charge} to support UAV charging. Vehicles that have close geographic location have a similar distance to the access point, thus, we can aggregate them into groups based on location, which we define them at the cell level. The set of cells in the region is defined as G. The parameters used are shown in TABLE I.

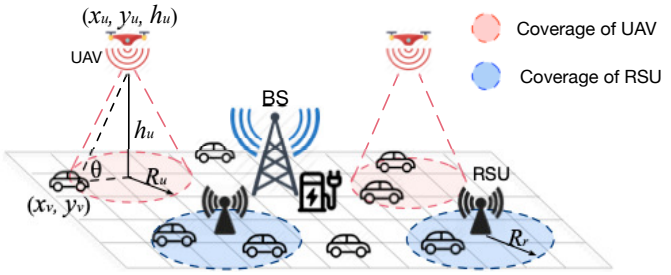


Fig. 1: Scenarios for UAV-assisted vehicular network

A. Channel Model

The signal-to-noise-ratio (SNR) between an infrastructure node i (BS, UAV or RSU) to a vehicle v can be expressed based on the average path loss defined in [19] as:

$$\eta_{i,v} = \frac{\rho_{i,v}}{10^{\overline{\varphi}_{i,v}/10}\sigma^2}, \tag{1}$$

where $\rho_{i,v}$ is the transmission power, $\overline{\varphi}_{i,v} \triangleq \mathbb{E}[\varphi_{i,v}]$ is the expected path loss, and σ^2 is the additive white Gaussian noise power at the receiver. We assume that the total bandwidth of an infrastructure node i is B_i and that this bandwidth is divided among the associated vehicles equally. The achievable rate of vehicle v (in bps) can be denoted as

$$x_v = \sum_{i \in I} \pi_{i,v} \frac{B_i}{N_i} log_2(1 + \eta_{i,v}), \tag{2}$$

TABLE I: List of notations used

Notation	Description
I	The set of all communication infrastructures.
S, R	The set of BSs and RSUs.
U	The set of UAVs, whose number is N.
V	The set of vehicles, whose number is $\|V\|$.
G	The set of cells in the region, whose number is $\|G\|$.
D	The vector of the number of vehicles in all cells.
$\mathcal{R}$	Coverage radius of a communication infrastructure.
l	Location of devices with coordinate (l^x, l^y, h), l^x and l^y are the horizontal coordinates and h is the elevation.
$\mathbf{L}$	The locations of UAVs.
$d_{i,j}$	Distance between two nodes i and j.
φ	The path loss of channel in dB.
X	The access rate vectors of vehicles $X = \{x_1, ..., x_{\|V\|}\}$.
x_v	Access rate of the vehicle v in bps.
$F^\alpha(X)$	The fairness criteria for vector X with α.
$\pi_{v,i}$	The binary accessing indicator to denote if the infrastructure node i offer service to the vehicle v.
$P(v)$	UAV energy consumption (J/s) with speed v.
$\Delta\tau$	Time interval between two steps.
ΔE_u	The energy consumption of UAV u.
$E_{u,t}$	The residual power of the UAV u at time t.
$T_{u,t}$	Time cost of UAVs for flying in time step t.

where $\pi_{i,v}$ is a binary association indicator, $\pi_{i,v} = 1$ denoting that infrastructure node i offers access service to vehicle v, and $\pi_{i,v} = 0$ otherwise; $N_i = \sum_{v \in V} \pi_{i,v}$ is the number of vehicles under service of the infrastructure node i. If a vehicle is in a range of more than one infrastructure node, we add the following constraint $\sum_{i \in I} \pi_{i,v} = 1$ to ensure that it can only be served by one infrastructure node.

Depending on the altitude of communication nodes, different channel models can be used to account for different propagation conditions. For UAV-assisted communication, it can be separated into two propagation slices, namely, ground-to-ground (G2G) and air-to-ground (A2G) [3].

1) Ground-to-ground Channels: G2G channel is below 10 m and 22.5 m for suburban and urban environments, respectively. G2G channels include communication between BSs, RSUs, and vehicles. Their large-scale channel attenuation (in dB) include distance-dependent path loss, whose classical model is log-distance path loss [2], [20] defined as follows:

$$\varphi_{i,v} = 10\zeta \log_{10}(d_{i,v}) + X_0 + X_\sigma, \tag{3}$$

where ζ is the path loss exponent that usually is in the range between 2 and 6; X_0 is the path loss at a reference distance with definition as $X_0 = 20\log(\frac{4\pi f_c d_0}{c})$, d_0 being the free-space reference distance, f_c is the carrier frequency and c is the light speed. $X_\sigma \sim \mathcal{N}(0, \sigma^2)$ is the shadowing effect, which is modeled as a normal (Gaussian) random variable with zero mean and a certain variance σ^2. Thus, the average path loss is equal to $\overline{\varphi}_{i,v} = 10\zeta \log_{10}(d_{i,v}) + X_0$.

2) Air-to-ground Channels: Obstructed A2G channel is in the range of 22.5-100 m for urban environments. Such a channel usually experiences a higher line-of-sight (LoS) probability than ground channels, however, it is still not 100%. In our scenario, UAVs communicate with ground vehicles in their coverage simultaneously by employing orthogonal frequency-division multiple access (OFDMA). For UAV-Vehicle (U2V) communication, the mmWave channel of 5G is used. The mmWave propagation channel of the U2V link is modeled using the standard log-normal shadowing model with LoS and non-line-of-sight (NLoS) links by choosing specific channel parameters [19]. In this model, the general path loss (in dB) between a UAV u and a vehicle v is defined as:

$$\varphi_{u,v} = \begin{cases} 10\zeta_{LoS}\log(d_{u,v}) + X_0 + X_{\sigma_{LoS}}, & \text{if LoS,} \\ 10\zeta_{NLoS}\log(d_{u,v}) + X_0 + X_{\sigma_{NLoS}}, & \text{if NLoS,} \end{cases} \tag{4}$$

where ζ_{LoS} and ζ_{NLoS} are the path loss exponents of LoS and NLoS links; $X_{\sigma_{LoS}}$ and $X_{\sigma_{NLoS}}$ are the shadowing random variables which are, respectively, represented as the Gaussian random variables (in dB) with zero means and σ_{LoS} and σ_{NLoS} as standard deviations. $X_0 = 20\log(\frac{4\pi f_m d_0}{c})$ is the reference path loss in the band of carrier frequency f_m of mmWave and at reference distance d_0.

The probability of having LoS transmission between UAVs and vehicles is expressed in [21] as:

$$P_{u,v}^{LoS} = \frac{1}{1 + ae^{-b(-\theta_{u,v}-a)}}, \ \forall i \in U, v \in V, \tag{5}$$

where a and b are constants that depend on the environment (e.g., the environment density, such as rural, urban); and θ is the elevation angle of a point on the ground with respect to a UAV (measured in degrees), can be expressed as $\theta_{u,v} = \frac{180}{\pi}\sin^{-1}(\frac{h_u}{d_{u,v}})$. Thus, the average path loss of A2G channels can be expressed as:

$$\overline{\varphi}_{u,v} = P_{u,v}^{LoS}\varphi_{u,v}^{LoS} + P_{u,v}^{NLoS}\varphi_{u,v}^{NLoS}, \tag{6}$$

where the probability of having NLoS is defined as $P_{u,v}^{NLoS} = 1 - P_{u,v}^{LoS}$. Besides, for high-altitude A2G (depending on the environment, e.g. in the range 100-300 m), the propagation is close to the free-space case. Thus, only LoS channel model is used, and $\overline{\varphi}_{u,v} = \varphi_{u,v}^{LoS}$.

B. Communications Coverage Model

We assume that RSUs are unmovable, thus their coverage radius is fixed and denoted as $\mathcal{R}_r$ for RSU r. For each UAV, given its height h_u, the path loss increases with the distance. Thus, the coverage of UAV u is defined as the maximum radius in which the path loss is below a value ($\overline{\varphi}_0$). The radius is defined as $\mathcal{R}_u = \sqrt{d_0^2 - h_u^2}$, where d_0 is the maximum distance between UAV u and the ground $d_0 = \arg\max_d(\overline{\varphi}_{u,v} \leq \overline{\varphi}_0)$.

We introduce Booleans to describe whether vehicles are in the coverage scope of infrastructure nodes or not, and they are defined as:

$$\forall\ i \in I,\ v \in V:\ e_{v,i} = \begin{cases} 1, & d_{i,v} \leqslant \mathcal{R}_i, \\ 0, & d_{i,v} > \mathcal{R}_i, \end{cases}$$

where $\mathcal{R}_i$ is the coverage radius of infrastructure node i. The set of candidate access points of vehicles is denoted as $M_v = \{i | e_{v,i} = 1, i \in I\}$. If the number of candidate access points is more than one, we assumed that the sequence of priority is UAV, RSU, and then BS. This means that if a vehicle is in the coverage of both a UAV and an RSU, it should choose the UAV; and only the vehicles that are out of any coverage

of RSUs and UAVs are controlled by the BS. If a vehicle is under two UAVs (or RSUs), it will choose the one with the shortest distance. Then, the definition of the access point of a vehicle v is given by:

$$m_v = \begin{cases} \arg\min_{u \in U} d_{v,u}, & \text{if } \sum_{u \in U} e_{v,u} > 0, \\ \arg\min_{r \in R} d_{v,r}, & \text{if } \sum_{u \in U} e_{v,u} = 0, \ \sum_{r \in R} e_{v,r} > 0, \\ S\ , & \text{otherwise.} \end{cases}$$

Thus, the factor $\pi_{i,v}$ in (2) is equal to 1 if $m_v = i$.

C. UAV Energy Consumption Model

Energy consumption for rotary-wing UAVs with speed v is modeled in [22], whose definition includes three items, namely, blade profile power, induced power, and parasite power. It physically represents the UAV energy consumption per second in Joule/second (J/s) with speed v:

$$P(v) = P_0(1 + \frac{3v^2}{v_0^2}) + P_1 \left(\sqrt{1 + \frac{v^4}{4v_1^4}} - \frac{v^2}{2v_1^2} \right)^{1/2} + \frac{1}{2} P_2 v^3, \quad (7)$$

where the P_0, P_1 and P_2 are coefficients of blade profile power, induced power, and parasite power, respectively. These coefficients are related to the aircraft's weight, air density, fuselage drag ratio, and rotor solidity, etc. v_0 and v_1 denote the tip speed of the rotor blade and the mean rotor induced velocity in hovering. We note that when the UAV is hovering with speed $v = 0$, the energy consumption is $P_h \triangleq P_0 + P_1$.

Combining both the propulsion energy and the communication related energy, the energy consumption of each UAV at time interval T can be expressed as:

$$\Delta E_u = \int_0^T \left(P(v_t) + P_c \sum_{i \in V_t} \pi_{u,i,t} \right) dt, \ \forall u \in U. \quad (8)$$

We assume that the communication related power is a constant in Watt (W), which is denoted as P_c. V_t is the set of vehicles at time t.

D. Spatial and Temporal Distribution of Vehicles

As mentioned before, we aggregate vehicles that have close geographic locations into groups, which we call them cells. To describe the characteristics of vehicles' distribution, we defined several factors to show their spatial and temporal variation in cells.

1) Spatial Variation: This factor can be used to show the variability in the number of vehicles between the cells. It is defined as the coefficient of variation, which is expressed as the ratio of the standard deviation σ to the mean μ as

$$SV(D) = \frac{\sigma(D)}{\mu(D)}, \quad (9)$$

where D is the set of the number of vehicles in all cells denoted as $D = \{|V_1|, |V_2|, ...|V_{|G|}|\}$, and $|V_i|$ is the number of vehicles in the cell i.

2) Temporal Variation: For each cell, we define another two factors to describe the temporal variation in the number of vehicles. First, to normalize the number of vehicles per cell, we define the proportion of the number of vehicles as

$$\forall g \in G, t \in [0, T] : \ \xi_{g,t} = \frac{|V_{g,t}|}{\sum_{i \in G} |V_{i,t}|}\ ,$$

where $|V_{g,t}|$ is the number of vehicles in the cell g at time step t, and G the total set of cells. Then, the time coefficient of variation (TCV) is proposed to describe the variation on vehicles' distribution over time from 0 to T. The TCV is the mean coefficient of variation of the time varying number of vehicles over all cells in G,

$$TCV(\xi_g) = \frac{1}{|G|} \sum_{g \in G} \frac{\sigma(\xi_g)}{\mu(\xi_g)}, \quad (10)$$

where ξ_g is the vector of proportional number of vehicles per cell g over the time period, which is a defined as $\xi_g = \{\xi_{g,t} | t \in [0, T]\}$. Besides, to describe the variation on each cell over adjacent time steps, the step coefficient of variation (SCV) is proposed. It is the average change in ξ_g between two adjacent time steps over all cells g, which is defined as

$$SCV(\xi_g) = \frac{\sum_{g \in G} \sum_{t=1}^{T} |\xi_{g,t} - \xi_{g,t-1}|}{|G|T}. \quad (11)$$

IV. Problem Formulation

A. Model of Placement Optimization

The problem of UAVs' dynamic placement can be formulated as follows: given a limited region in which some terrestrial communication infrastructures have been deployed, we try to find the optimal placement of a limited number of UAVs N with an objective function calculated over the bandwidth allocated to vehicles. We want our objective function to account for the fairness of the allocation of the aggregate network resources, both UAVs and terrestrial. For that, we consider that at the beginning ($t = 0$) all of the UAVs are located at the charge station. The optimization of UAVs dynamic placement with the variable $\mathbf{L}$ can be modeled as follows:

$$\max_{\mathbf{L}} \int_0^T \frac{F^\alpha(X_t)}{|V_t|} dt$$

$$\text{s.t.} \ \ \forall u \in U, t \in T :$$

$$l^x_{min} \le l^x_{u,t} \le l^x_{max}, \quad (12)$$

$$l^y_{min} \le l^y_{u,t} \le l^y_{max}, \quad (13)$$

$$h_{min} \le h_{u,t} \le h_{max}, \quad (14)$$

$$E_{u,t} > 0. \quad (15)$$

This problem has three main constraints. Firstly, UAVs' locations should be within the limited region and between the lower-bound l_{min} and the upper-bound l_{max} with constraints of equations (12), (13) and (14). Secondly, the UAVs shouldn't be out of energy, and they can go back for charging with variables to be $l_u = l^{charge}$. Depending on whether UAVs'

target position is a charge station or not, we can class the UAVs into two categories: working UAVs and charging UAVs.

The objective of the maximization is a sum over time, normalized by the time-varying number of vehicles, of a parametric function modeling the fairness and efficiency of the allocation of network resources, UAVs, and terrestrial. To account for as many bandwidth allocation scenarios as possible, we use a parametric fairness function that is widely known in the literature on resource allocation on the Internet and TCP congestion control. Given a positive constant α and access rate vectors of vehicles $X = \{x_1, ..., x_{|V|}\}$, this fairness function is given in [13] as

$$F^{\alpha}(X) = \begin{cases} \frac{1}{1-\alpha} \sum_{v=1}^{|V|} x_v^{1-\alpha}, & \text{if } \alpha \neq 1, \\ \sum_{v=1}^{|V|} \log x_v, & \text{otherwise.} \end{cases} \tag{16}$$

This function corresponds to the sum of vehicles' access rates (namely total throughput) as $\alpha \to 0$; the proportional fairness when $\alpha \to 1$; the harmonic mean fairness when $\alpha \to 2$; the max-min fairness when $\alpha \to \infty$, which tries to maximize the minimum access rate overall vehicles by giving priority to vehicles that have achieved lowest data rate. Proportional fair and harmonic mean fair allocations are trade-offs between maximizing total throughput and max-min fair allocations. The first scenario of $\alpha = 0$ is the mere scenario where the throughput is maximized independently of any fairness consideration.

B. Time Discretization for the Problem

An intuitive way to maximize fairness is to select the optimal locations at each time interval of $\Delta\tau$. The time discretization variables are $\mathbf{L} = \{L_t | t \in [1, T]\}$, and $L_t = \{l_{u,t} | u \in U\}$. The UAVs are assumed to have two states in each time interval, namely, flying to the target locations and hovering to serve the incoming flows, or flying back to the charge station and charging. During the flying period, we assume that the UAVs have a constant velocity of v_m. The time cost by UAV u for flying from the current location $l_{u,t-1}$ to the target location $l_{u,t}$, which may be different from each other, is denoted as:

$$T_{u,t} = \frac{||l_{u,t} - l_{u,t-1}||}{v_m}, \ \forall u \in U.$$

Based on the flying time, the ascending order of the UAVs' arriving to offer service of time step t is defined as $\mathcal{I}_t$. ΔT refers to the time interval between two consecutively arrivals of UAVs as shown in Fig. 2. The time interval of i-th arrivals in time slot t can be expressed as:

$$\Delta T_{i,t} = \begin{cases} \Delta\tau - T_{u_i,t}, & \text{if } u_i \text{ is last to arrive} \\ T_{u_{i+1},t} - T_{u_i,t}, & \text{otherwise} \end{cases}. \tag{17}$$

where $T_{u_i,t}$ is the time cost for flying of the i-th arrival UAV u_i. During $\Delta T_{i,t}$, the first i arriving UAVs can offer services. Especially, when the last one arrived at the target location, all the working UAVs will offer services till the end of this time slot $\Delta\tau$. In addition, if a UAV with $T_{u,t} \geq \Delta\tau$, it won't have time to work as the access points in this time slot.

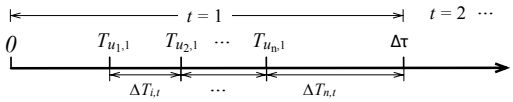


Fig. 2: Time interval between the arrival of UAVs

During the hovering period, the UAVs hover at their target locations and serve the incoming flows. The mean fairness value of each step is defined as

$$f_t^{\alpha} = \frac{\sum_{i \in \mathcal{I}_t} F^{\alpha}(X_{t,i}) \Delta T_{i,t}}{|V_t| \Delta\tau}, \tag{18}$$

where $X_{t,i}$ is the vector of access rate of vehicles when the i-th UAV arrived to offer services in time t. $X_{t,i}$ is calculated based on the mean distribution of vehicles over the time slot t, and $|V_t|$ is the mean number of vehicles. The objective function can be approximated as $\sum_{t=0}^{T} f_t^{\alpha}$.

To make sure each UAV is not out of energy, we assume that when its energy $E_{u,t}$ is less than some level $\tilde{E}$ (e.g. 10%), it needs to go back to the charging station. We used Boolean parameters ϕ_u to denote whether UAV u is in service (with value 1), or out of service to charge (with value 0).

$$\forall\ u \in U : \phi_{u,t} = \begin{cases} 0, & \text{if } E_{u,t} \leqslant \tilde{E} \\ 1, & \text{otherwise} \end{cases}. \tag{19}$$

Equation (8) can be reformulated as:

$$\Delta E_{u,t} = P(v_m) T_{u,t} + \phi_{u,t} (P_h + P_c N_{u,t}) (\Delta\tau - T_{u,t}), \tag{20}$$

where $N_{u,t}$ is the number of vehicles that are connected with UAV u with definition as $N_{u,t} = \sum_{i \in V_t} \pi_{i,t}^{u}$. The first item is the flying power, the second item is the hovering and service power. If the UAVs are selected to go back to charge with $\phi_{u,t} = 0$, there would be no power cost for hovering and offer services. The updating of the battery of UAVs is expressed as:

$$\forall\ u \in U : E_{u,t+1} = \begin{cases} E_{u,t} - \Delta E_{u,t}, & \text{if } \phi_{u,t} = 1 \\ E^{full}, & \text{if } \phi_{u,t} = 0 \end{cases}, \tag{21}$$

where E^{full} denotes the capacity of UAVs' battery.

V. DRL for UAVs' Dynamic Placement

In this section, we introduce the proposed DRL-based method to solve the UAVs' dynamic placement problem.

A. DRL-based Problem Description

We model trial-and-error learning as the Markov decision process (MDP). At each time t, the agent observes the current state s_t of the interactive environment and gives an action a_t according to its policy. Then, the environment returns reward r_t as feedback, and translates to the next state s_{t+1} according to the transition probability $P(s_{t+1} | s_t, a)$. The goal to find an optimal policy can be formulated as the mathematical problem of maximizing the expectation of cumulative discounted return $R_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$, where $\gamma \in [0, 1]$ is a discount factor for future rewards to dampen the effect of future rewards on the

action; r_k is the reward of each step, and T is the time horizon before game over.

In our scenario, we assumed that the agent is deployed in a BS. The vehicles and UAVs in the managed region send their states (e.g. geographic locations) to the agent through the network periodically, e.g. every 2 minutes. Thus, the agent can obtain the vehicular traffic of the environment as well as the state of UAVs. Then, the agent makes decisions on UAVs' next locations considering the states. Next, the agent sends the decisions back to UAVs. Noticed that the UAVs' locations can't be out of communication coverage, thus, there are some constraints on the actions. The UAVs follow instructions from the agent and fly to their target locations. The key definitions of DRL for UAV placement are expressed as follows.

State Space: The state is defined as $s_t = \{D_t, L_t, E_t\}$ of each time step t. Firstly, the state should include the distribution of vehicles D_t, which is a vector of the mean number of vehicles in all cells over the time interval of $\Delta\tau$. Secondly, the current locations of UAVs would affect the decisions on the next locations, thus, the vector of current UAVs' locations $L_t = \{l_{u,t} | u \in U\}$ is used as the input. Then, the last item is the vector of UAVs' residual energy $E_t = \{E_{u,t} | u \in U\}$, which can also affect the decisions on next locations. For example. UAVs may prefer to fly nearby to save energy and offer service for a longer time when the reward is inadequate to let them fly far away.

Action Space: The action is defined as a vector of UAVs' normalized locations $a_t = \{w_{u_1,t}, ..., w_{u_N,t}\}$ of each time step t, in which $w_{u,t} = \{w^x_{u,t}, w^y_{u,t}, w^h_{u,t}\}$ is a vector denoted the normalized 3D coordinates of the UAV u. The next location of the UAV u can be calculated with $l^x_{u,t} = w^x_{u,t}(l^x_{max} - l^x_{min}) + l^x_{min}$, and for the other two coordinates are similar. In this way, the UAVs' locations are limited to some a certain range. Besides, the UAVs' next locations depend on not only the action a_t but the charging factor ϕ_t. Then, the next location of UAV u is defined as:

$$\forall\ u \in U : l_{u,t} = \begin{cases} l_{u,t}, & \text{if } \phi_{u,t} = 1 \\ l^{charge}, & \text{if } \phi_{u,t} = 0 \end{cases}. \tag{22}$$

Reward: The value of reward depends both on the current state and on the taken action. In this problem, the reward is defined as the increment ratio of fairness value of vehicles' accessing rate to the original fairness value without UAVs working for communication, which is

$$r_t = \frac{(f^\alpha_t - f^\alpha_t(0))}{|f^\alpha_t(0)|} \tag{23}$$

where f^α_t is expressed in (18), and $f^\alpha_t(0)$ is the original fairness value without UAVs working for communication.

B. DRL-based Training Algorithm

Deep Deterministic Policy Gradient (DDPG) [23] is used as the training algorithm for UAVs' placement shown in Fig. 3.

Actor: The actor is responsible for making actions a based on the observed states s. The actor is designed to be a neural network, which takes continuous policies μ with regard to parameters θ^μ. As for deterministic policies, we define actions as $a = \mu(s|\theta^\mu)$.

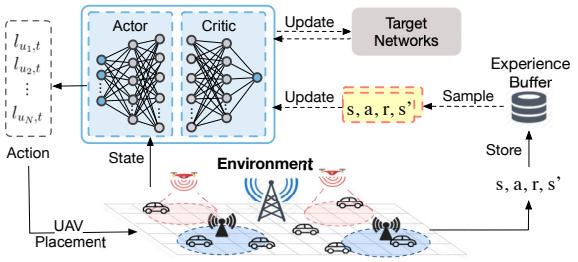


Fig. 3: DDPG for UAV dynamic placement.

Critic: The critic is fed with states s and actions a. The critic $Q(s, a|\theta^Q)$ is defined as a action-value function for agent, whose parameters are θ^Q.

Target Networks: The target policy μ' is with parameters $\theta^{\mu'}$, and the target critic Q' is with parameters $\theta^{Q'}$. These parameters are periodically updated as:

$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}, \theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \tag{24}$$

where τ is a coefficient between 0 and 1.

Experience Replay Buffer: For off-policy training, the experience replay buffer, denoted as $\mathcal{M}$, is used to store experiences and offer samplings for training. It is in the form of tuples, denoted as (s, a, r, s') with states, actions, rewards, and successor states of the next time.

Based on the introduction of key components of DDPG, we will explain the main process of DDPG with Fig. 3. In each execution, first, to explore the action space, random noise is added to the output of the actor-network $a = \mu(s|\theta^\mu) + \mathcal{N}$, where $\mathcal{N}$ is the exploration noise. Then, UAVs fly to the next location and obtain the rewards r and new state s'. Next, the experience (s, a, r, s') is stored in the experience replay buffer. In each training episode, a batch of transitions is randomly sampled for off-policy training. The parameters θ^μ of the actor-network are updated using the sampled policy gradient:

$$\begin{aligned} \nabla_{\theta^\mu} J(\mu) &= \mathbb{E}_{s\sim\mathcal{M}}[\nabla_{\theta^\mu} Q(s, \mu(s))] \\ &= \mathbb{E}_{s\sim\mathcal{M}}[\nabla_{\theta^\mu}\mu(s)\ \nabla_a Q(s, \mu(s))], \end{aligned} \tag{25}$$

where the symbol $\mathbb{E}\{\cdot\}$ denotes the expectation value, and states s are sampled from experience replay buffer $\mathcal{M}$. The parameters θ^Q of the critic network are updated to minimize the loss:

$$\mathcal{L}(\theta^Q) = \mathbb{E}_{s,a,r,s'\sim\mathcal{M}}[Q(s, a) - (r + \gamma Q'(s', \mu'(s')))]^2, \tag{26}$$

where the 4-tuples (s, a, r, s') are sampled from $\mathcal{M}$, and a' is obtained with function $a' = \mu'(s')$ of the target actor.

C. Overhead Analysis

Although DRL is well-suited to tackle problems with longer-term planning using high- dimensional observation, it brings in more time to collect data and training, as well as costs more memory resource for storing experiences. Firstly, we analyze the computational complexity of the proposed DRL.

The deep neural network can be viewed as matrix multiplication. Thus, the complexity for the actor is approximated to $O(|G|HN)$, where $|G|$ is the number of cells, H is the number of hidden layers, and N stands for the number of the UAVs. Similarly, the complexity of the critic neural network is approximated to $O(|G|HN)$. The training procedure of DDPG has approximate complexity $O(N_{ep}K_sNH|G|)$, where N_{ep} is the number of training episodes, and K_s is the batch size. Thus, the training is affected by the number of UAVs, which means using more UAVs will increase complexity. For the real-time execution, namely giving UAVs' locations based on a state, the complexity is approximated to $O(|G|HN)$. However, if the brute force method is used to find optimal locations, which choose locations with the highest reward from K possible locations, the complexity is $O(K^N)$. Noticed that this method has high complexity. Secondly, the memory size depends on the maximum number of transitions that can be stored in the experience replay buffer, denoted as $|\mathcal{M}|$. All data is assumed in 32-bit float, thus, each state is $4(|G|+4N)$ Bytes, each action is $12N$ Bytes, and each reward is 4 Bytes. Then, one transition is $8|G|+44N+4$ Bytes, and the total size of memory for the experience reply buffer is $|\mathcal{M}|(8|G|+44N+4)$ Bytes.

VI. Experiment and Results

We conduct the simulation to evaluate the performance of the proposed DRL-based UAV dynamic placement. In this section, we first introduce the evaluation settings and then present results and analysis. Our code, as well as the data used in our simulation, are uploaded to GitHub [24].

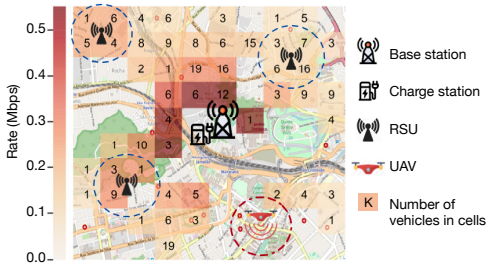


Fig. 4: Map of Rio de Janeiro for experiments.

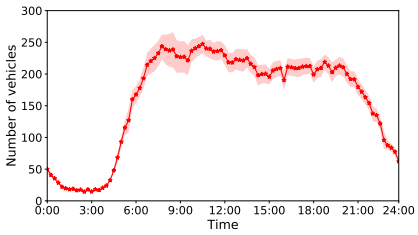


Fig. 5: Number of Vehicles in 3×3 km^{2}of Rio.

A. Evaluation Settings

Fig. 4 shows the region used as the geographical footprint of the experiments. This region consists of 3×3 km^2 area of Rio de Janeiro, Brazil, with a BS and a charging station located at the center, and 3 RSUs. The BS has a coverage radius of around 2 km, and each RSU has a 300 m coverage radius. For UAVs, the coverage radius depends on its height, which is in the range of 50 m to 100 m. Besides, the map is cut to 10*10 cells, thus, the number of cells $|G|$ is 100. The replay buffer $|\mathcal{M}|$ is set to be 50000, thus, the size of memory is around 50 MB.

TABLE II: List of notations used

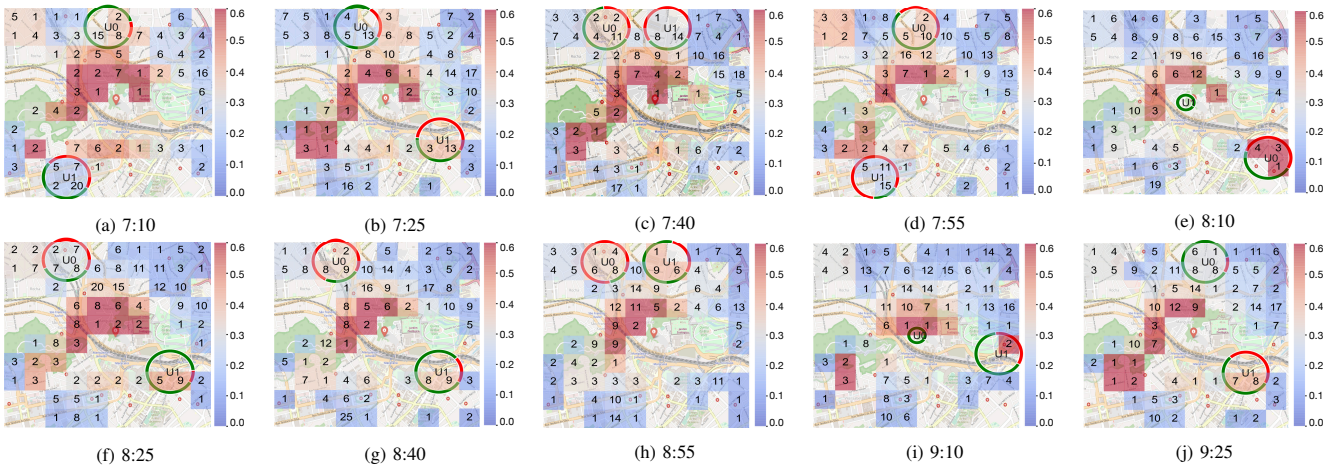
Notation	Description	Value
f_c, f_m	The carrier frequency of cellular and mmWave	2, 38 Ghz
B_i	Total bandwidth of BSs, UAVs and RSUs	60, 1, 1Mhz
v_0	The tip speed of the rotor blade	120 m/s
v_1	Mean rotor induced velocity in hover	4.03 m/s
v_m	The fix flying speed of UAVs	20 m/s
P_0	The coefficient of blade profile power	79.85 J/s
P_1	The coefficient of induced power	88.63 J/s
P_2	The coefficient of parasite power	0.018 kg/m
P_c	The communication power of UAVs	1 W
$\overline{\varphi}_0$	The path loss rate limitation of UAVs	-138 dBm
a, b	The constants of (5)	10, 0.6
ζ	The path loss exponent of LoS and NLoS	2, 2.4
σ^2	The additive white Gaussian noise in (1)	-95 dBm
E^{full}	The capacity of UAVs' battery	700 kwh

For vehicle mobility, and to be as realist as possible, we use a dataset available in [25]. This dataset offers real-time position data reported by buses from the city of Rio de Janeiro (24h format). In this experiment, we select vehicle data of one week i.e., $\varpi = 7$, and we consider every 15 minutes to be one step $\Delta\tau = 15$, which means that the agent takes actions every 15 minutes. Thus, for one day, there are 96 steps in total ($T = 96$). Fig. 5 shows the number of vehicles $|V_t|$ with the time t. The curve shows the 68% confidence interval of the number of vehicles for 7 days. We analyze the characters of vehicles' distribution to assess traffic dynamic and possible benefits of UAVs to improve communication. For the dataset, the mean of SV over 7 days is 195%, which is defined as $\frac{\sum_{d=1}^{\varpi}\sum_{t=1}^{T} SV(D_{d,t})}{T\,\varpi}$ where SV is in (9), and $D_{d,t}$ is the set of the number of vehicles in all cells at time t in day d. The mean of TCV defined in (10) and SCV in (11) over 7 days are 152.4% and 43.5%, respectively. These values show that the distribution of vehicles varies greatly over time and between cells. It should be possible to accommodate this variability by placing UAVs differently and therefore improve the allocation of resources according to the predetermined fairness objective. Simulation parameters and their values are listed in TABLE II.

B. Results and Performance Evaluation

We first show some visualizing results of UAVs' optimal and dynamic placement in Fig. 6. It shows results from 7:00 to 9:00 with 2 UAVs and $\alpha = 0$. In this figure, the residual battery (green part) and used battery (red part) of UAVs are also shown. Notice that U1 is lower than 10% in Fig. 6e and consequently goes back to the charging station.

Performance with learning. We show the performance of our proposed method in Fig. 7 with $\alpha = 1$ and different numbers of UAVs ($N = 1, 2, 3, 4$). The horizontal axis of this figure is the learning episodes. The vertical axis represents the mean value of accumulated rewards $\frac{1}{\varpi}\sum_{d=1}^{\varpi}\sum_{t=1}^{T} r_t$. The curves show the



(a) 7:10 (b) 7:25 (c) 7:40 (d) 7:55 (e) 8:10

(f) 8:25 (g) 8:40 (h) 8:55 (i) 9:10 (j) 9:25

Fig. 6: Locations and battery left of UAVs with $\alpha = 0$ and $N = 2$ from 7:00 to 9:30.

68% confidence interval of reward for 7 days. We can conclude that our proposed method using DDPG has good convergence. Besides, with more UAVs, the algorithm needs more episodes to converge, since the learning is more challenging with the increasing dimension of states and actions. Similar results can also be obtained with other α values.

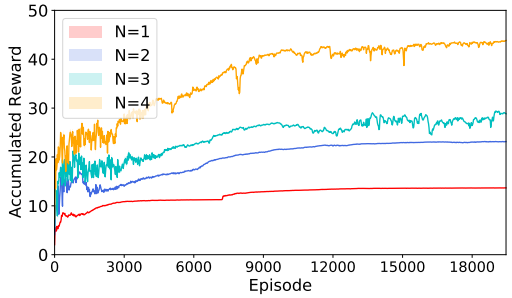


Fig. 7: Accumulated reward with training with $\alpha = 1$.

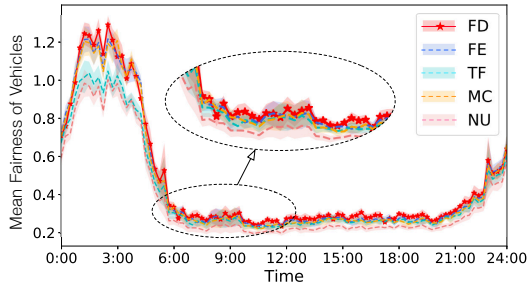


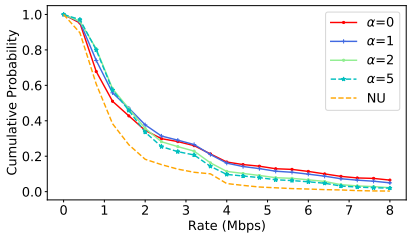
Fig. 8: Mean fairness value of vehicles with $\alpha = 1$ and $N = 2$.

Performance compared with baselines. We show simulation results for our proposed assignment methodology (called **FD** for short) and compare them with three other approaches: (1) Fairness-based placement using enumerations (called **FE** for short), which choose locations with the highest reward from $K = 100$ possible locations at each step. It is used to show how good the DRL is compared with the brute force method. (2) Maximum coverage placement [9] (called **MC** for short) using DRL. Although it can cover vehicles as many as possible by UAVs, it ignores the difference between vehicles since some of them already have been good served. It is used as a comparison since we want to show the importance of cooperation with terrestrial infrastructures. (3) Coverage time fairness [12] based placement using DRL (called **TF** for short). It considers the fairness in coverage time but also ignores the difference between covered areas due to the uneven traffic and terrestrial infrastructures' locations. (4) No UAVs (called **NU**).

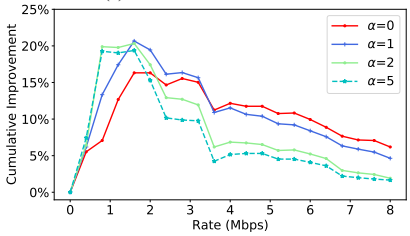
We show the performance of mean fairness value of vehicles $\frac{F^{\alpha}(X_t)}{|V_t|}$ at different times with $\alpha = 1$ and $N = 2$ compared with the baselines. The curves show the 68% confidence interval for 7 days. Combined with Fig. 5, we can see that the mean fairness decreases with the increasing number of vehicles. For example, during the rush hour between 6:00 and 21:00, the means fairness is around 0.3, and between 2:00 to 4:00, it is much higher at around 1.2. Besides, comparing with FE, MC, TF, and NU, the mean improvement of FM is around 4.62%, 6.60%, 11.08%, and 23.88%, respectively.

Performance with different fairness factors. Fig. 9 shows the cumulative distribution of vehicles' access rates in Fig. 9a and Fig. 9c, as well as the improvement in terms of the cumulative distribution of vehicles' access rates compared with NU in Fig. 9b and Fig. 9d. These curves show performance with 2 UAVs and different α. Fig. 9a illustrates the cumulative probability from 1:00 am to 5:00 am in 7 days, whose rate is mostly between 0.5 to 8 Mbps. Fig. 9c represents the rush hour from 9:00 am to 13:00 am in 7 days, whose rate is mostly between 0.2 to 1 Mbps. Seen from Fig. 9b and Fig. 9d, compared with NU, all curves with different α values have a significant improvement. Comparing different α values, we can conclude that UAVs are more likely to serve vehicles at a lower rate with a higher α value.

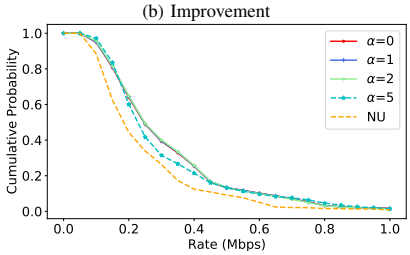
Fig. 10 shows the total throughput defined as the sum of vehicles' access rates with different values of α and different numbers of UAVs. The lines show the interval for 7 days. We observe that when more UAVs are deployed in the same environment, the overall throughput increases. The total throughput grows with decreasing α, since lower α more



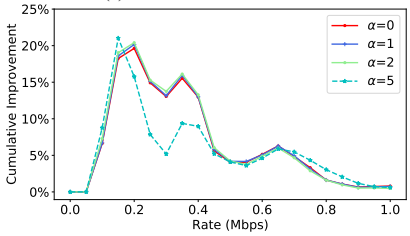
(a) Cumulative distribution



(b) Improvement



(c) Cumulative distribution



(d) Improvement

Fig. 9: Cumulative distribution and improvement on cumulative distribution of vehicles' access rate with $N = 2$ from 1:00 to 5:00 (a)-(b) and from 9:00 to 13:00 (c)-(d).

inclines to improve the overall performance. In particular, $\alpha = 0$ has the highest total throughput as expected since the objective of this allocation policy is true to maximize the total throughput.

Performance with different numbers of UAVs. Fig. 11 shows the cumulative distribution of vehicles access rate for 7 days when $\alpha = 1$. The four curves are with different numbers of UAVs deployed (from 1 to 4). The more UAV deployed the higher rate can be obtained. For example, around 27%, 35%, 40%, and 45% of vehicles have more than 0.4 Mbps with one to four UAVs, respectively. With the other α values, similar results can be observed.

In addition, Fig. 12 shows the access rate in time latitude with different numbers of UAVs. Fig. 12a and Fig. 12b shows the percentage of vehicles whose access rate is more than 1 Mbps and 0.5 Mbps. First, the more UAVs are deployed, the higher the access rate, whether during rush hour or not.

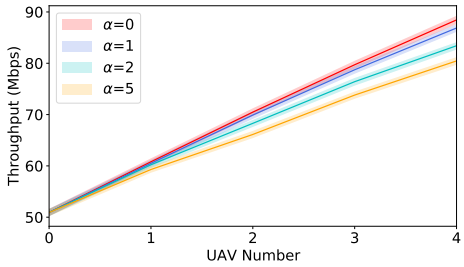


Fig. 10: Throughput with different numbers of UAVs and α.

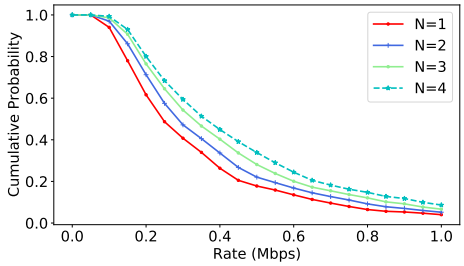
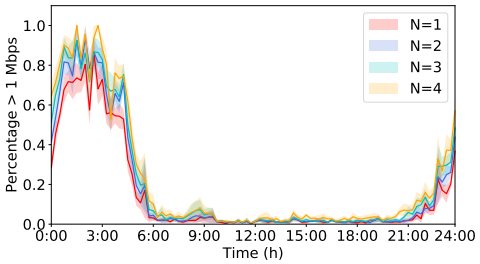
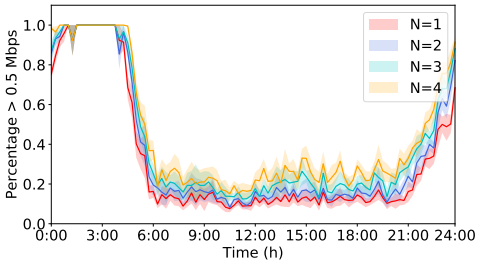


Fig. 11: Cumulative distribution of vehicles' access rate with $\alpha = 1$.

Second, at the peak hours from 6:00 to 21:00, the improvement by increasing the number of UAVs is more evident at a lower rate, e.g., 0.5 Mbps. For other times of the day, the improvement is more obvious for the upper part, e.g., 1 Mbps.



(a) Percentage of vehicles whose access rate is more than 1 Mbps



(b) Percentage of vehicles whose access rate is more than 0.5 Mbps

Fig. 12: Performance with different numbers of UAVS

VII. Conclusion

In this paper, we leverage multiple 5G UAVs to enhance network resource allocation among vehicles by dynamic placement on-demand. We propose a DRL approach to determine the position of UAVs to improve the fairness and efficiency of

network resource allocation while considering the UAVs' flying range, communication range, and limited energy resources. We use a parametric fairness function for resource allocation that can be tuned to reach different allocation objectives ranging from maximizing the total throughput of vehicles, maximizing minimum throughput, as well as achieving proportional bandwidth allocation. The results of our simulations show that the dynamic placement of UAVs can improve the fairness of communication. Besides, the UAVs' locations are affected by the setting of fairness criteria, since they have different preference to serve different vehicles. Furthermore, we demonstrate that by deploying more UAVs in the same environment it is possible to improve fairness by serving more vehicles poorly connected, but with the cost of increasing the training time due to increasing computational complexity.

ACKNOWLEDGMENT

This work was partly funded by Inria, supported by the French ANR "Investments for the Future" Program reference #ANR-11-LABX-0031-01, and UNICAMP, through the FAPESP Grant number #2017/50361-0, both in the context of the DrIVE #EQA-041801 associated team.

REFERENCES

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1123–1152, 2015.

[2] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *arXiv preprint arXiv:1903.05289*, 2019.

[3] E. Vinogradov, H. Sallouha, S. De Bast, M. M. Azari, and S. Pollin, "Tutorial on UAV: A blue sky view on wireless communication," *arXiv preprint arXiv:1901.02306*, 2019.

[4] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, 2017.

[5] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3417–3442, 2019.

[6] R. Ghanavi, E. Kalantari, M. Sabbaghian, H. Yanikomeroglu, and A. Yongacoglu, "Efficient 3d aerial base station placement considering users mobility by reinforcement learning," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[7] J. Chen and D. Gesbert, "Optimal positioning of flying relays for wireless networks: A los map approach," in *2017 IEEE international conference on communications (ICC)*. IEEE, 2017, pp. 1–6.

[8] V. Saxena, J. Jaldén, and H. Klessig, "Optimal UAV base station trajectories using flow-level models for reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 1101–1112, 2019.

[9] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different qos requirements," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 38–41, 2017.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *IEEE SIGNAL PROCESSING MAGAZINE*, 2017.

[12] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, 2018.

[13] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on networking*, vol. 8, no. 5, pp. 556–567, 2000.

[14] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of uav-mounted mobile base stations," *IEEE Communications Letters*, vol. 21, no. 3, pp. 604–607, 2016.

[15] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1647–1650, 2016.

[16] P. S. Bithas, E. T. Michailidis, N. Nomikos, D. Vouyioukas, and A. G. Kanatas, "A survey on machine-learning techniques for UAV-based communications," *Sensors*, vol. 19, no. 23, p. 5170, 2019.

[17] T. Yuan, W. B. da Rocha Neto, C. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, "Harnessing machine learning for next-generation intelligent transportation systems: A survey," 2019.

[18] M. S. Shokry, D. Ebrahimi, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Leveraging UAVs for coverage in cell-free vehicular networks: A deep reinforcement learning approach," *IEEE Transactions on Mobile Computing*, 2020.

[19] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1046–1061, 2017.

[20] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[21] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[22] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, 2019.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[24] https://github.com/TingtingYuan/UAV_fairness.git.

[25] D. Dias and L. H. M. K. Costa, "CRAWDAD dataset coppe-ufrj/riobuses (v. 2018-03-19)," Downloaded from https://crawdad.org/coppe-ufrj/RioBuses/20180319, Mar. 2018.