Federated Multi-Agent Deep Reinforcement Learning for Dynamic and Flexible 3D Operation of 5G Multi-MAP Networks

Esteban Catté, Mohamed Sana and Mickael Maman
CEA-Leti, Universite Grenoble Alpes, F-38000 Grenoble, France
{esteban.catte, mohamed.sana, mickael.maman}@cea.fr

Abstract—**This paper addresses the efficient management of Mobile Access Points (MAPs), which are Unmanned Aerial Vehicles (UAV), in 5G networks. We propose a two-level hierarchical architecture, which dynamically reconfigures the network while considering Integrated Access-Backhaul (IAB) constraints. The high-layer decision process determines the number of MAPs through consensus, and we develop a joint optimization process to account for co-dependence in network self-management. In the low-layer, MAPs manage their placement using a double-attention based Deep Reinforcement Learning (DRL) model that encourages cooperation without retraining. To improve generalization and reduce complexity, we propose a federated mechanism for training and sharing one placement model for every MAP in the low-layer. Additionally, we jointly optimize the placement and backhaul connectivity of MAPs using a multi-objective reward function, considering the impact of varying MAP placement on wireless backhaul connectivity.**

Index Terms—**Mobile Access Points, Integrated access backhaul, Multi-agent Deep Reinforcement Learning, Federated Learning, mmWave Communications, Dynamic 5G Networks.**

I. INTRODUCTION

5G aims to offer fair opportunities for User Equipments (UE) regardless of their location or mobility via efficient management. Mobile Access Points (MAPs), which are Unmanned Aerial Vehicles (UAV), are gaining attention as a flexible infrastructure, useful for various applications [1]. MAPs can collaborate to form a Multi-MAP network, but there is limited research on managing them in dynamic networks with user mobility, interference, varying traffic, and fluctuating MAP numbers. Our objective is to efficiently manage multiple MAPs in terms of their number, placement, and trajectory while considering dynamic constraints over a longer time scale than the current state-of-the-art approaches. Previous studies have explored different approaches leveraging the 3-dimensional (3D) mobility of MAPs, but often without accounting for all the dynamic network constraints simultaneously. For instance, in [2], the authors proposed an iterative optimization method for MAP placement based on user mobility. Another study by Ghanavi et al. [3] extended the scenario to multiple MAPs managed by a reinforcement Q-learning algorithm. Wang et al. [4] introduced a virtual forces algorithm based on statistical user distributions for computing network cartography. It is worth noting that user distribution can impact MAP numbers and deployment positions, even when the number of UEs remains constant. These diverse solutions demonstrate the variety of MAP management techniques, highlighting the need for iterative approaches to efficiently handle dynamic network constraints. However, ensuring long-term performance in a constantly changing network remains a challenge.

The aforementioned papers highlight the potential of using a greedy MAPs deployment approach to determine their optimal number. For instance, in [5], [6], [7], [8], [9], proposed solutions adjust the number of deployed MAPs iteratively to meet network constraints. However, this approach may suffer from convergence delays and does not account for network evolution. In contrast, our study proposes a hierarchical architecture that dynamically determines the number of MAPs for user coverage, independent of the placement procedure. Our architecture aims to strike a balance between cost and coverage by determining both the number and positions of MAPs, as these aspects affects each other.

Obviously, MAP management must adapt to changing network conditions, including trajectory adjustments. In [10], authors used a successive convex optimization to optimize MAP trajectories and UE data rates under mobility constraints. However, a significant breakthrough in MAP trajectory optimization has been achieved with Multi-Agent Deep Reinforcement Learning (MADRL) models. In [11] and [12], authors proposed target MADRL models based on the actor-critic architecture to handle multiple factors. Authors of [13] proposed a MADRL approach with pre-deployed MAPs on UE clusters. This approach takes advantage of the low-complexity deployment algorithm and the ability of MADRL model to adjust positions in complex environments.

Our paper presents a problem formulation and proposes a two-level hierarchical architecture based on joint optimization for a dynamic 5G network while considering Integrated Access-Backhaul (IAB) constraints. The decision process is scalable and distributed and it determines the number of MAPs through consensus in the high-layer. In the low-layer, MAPs manage their placement using our previously proposed dual-attention based DRL model [14] that encourages cooperation without any a-priory information or retraining procedure. To increase the generalization ability of learned model, reduce complexity and improve performance in novel scenarios, we propose a federated mechanism that involves training and sharing one placement model for every MAP, as suggested in [15].

Additionally, we aim to jointly optimize backhaul connectivity of MAPs using a multi-objective reward function, considering the impact of varying MAP placement on wireless backhaul link as highlighted in previous studies [16] and [17].
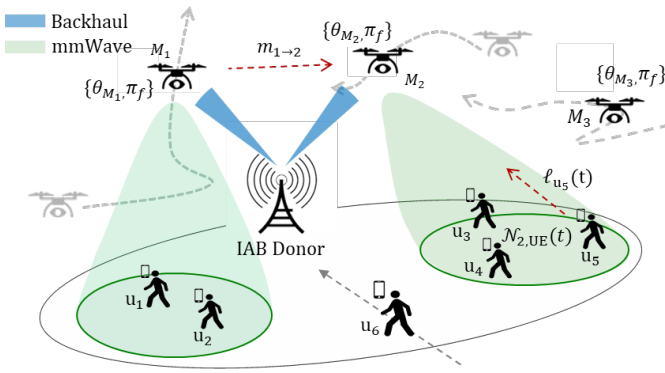
Fig. 1. System model with one IAB Donor, two deployed MAPs maintaining their trade-off value $\{\theta_{M_1}, \theta_{M_2}\}$ and sharing policy π_f with one joining MAP, five communicating UEs with one joining UE and corresponding links.

The paper is organized as follows. Section II presents the system model and Section III formulates the addressed problem. Then, Section IV describes our proposed solution, whereas Section V provides our numerical results. Finally, Section VI concludes the paper.

II. SYSTEM MODEL

We consider a downlink network composed of M MAPs operating at mmWave frequencies. Each flying MAP can establish a backhaul link with a grounded IAB donor. We define $M_s(t)$ as the number of deployed MAPs at time t, which move to provide services to $K(t)$ UEs. Let $\mathcal{U}(t) = \{1, \ldots, K(t)\}$ be the set of UEs, $\mathcal{S}_0(t)$ the set of all Base-Station (BS) including the IAB donor indexed by 0 and $\mathcal{S}(t) = \{1, \ldots, M\}$ denotes the dynamic set of deployed MAPs. We assume that UEs can be associated with only one MAP $i \in \mathcal{S}_0(t)$ providing the maximum signal-to-noise ratio (via max-SNR algorithm). In our system model, we assume that the grounded location $\ell_j(t) \in \mathbb{R}^2$ of UEs changes with time, requiring dynamic and on-demand reconfiguration of MAPs deployment. Once deployed, MAP $i \in \mathcal{S}(t)$ can adapt its 3D location $\ell_i(t)$ in a region $\mathcal{L}$ of $\mathbb{R}^3$ space and can only serve at most $K_i(t)$ UEs due to limited beamforming capability. In this dynamic network, optimizing the number and placement of MAPs is a challenging and important task to improve network spectral efficiency. Indeed, MAPs should dynamically adjust their number and location to follow UE's dynamics while limiting interference.

A. Channel Modeling

Our system model considers an *out-of-band* relaying IAB network where the access and backhaul links are orthogonal and do not interfere on each other. In this context, we split the available mmWave bandwidth B into parts dedicated to backhaul (μB) and access network ($(1 - \mu)B$), where $\mu \in [0, 1]$. We assume that both the access and backhaul links use Spatial Division Multiple Access (SDMA). Thus, when UE j is receiving data from BS i, it experiences a downlink signal-to-interference-plus-noise ratio $\text{SINR}_{i,j}^{(a)}$, which reads as:

$$\text{SINR}_{i,j}^{(a)}(t) = \frac{\zeta_{i,j}(t) P_{i,j}^{\text{Tx}} G_{i,j}^{\text{Tx}}(t) G_{i,j}^{\text{H}}(t) G_{i,j}^{\text{Rx}}(t)}{I_{i,j}^{(a)}(t) + (1 - \mu) N_0 B}. \quad (1)$$

Here, $P_{i,j}^{\text{Tx}}$ is the transmit power from BS i towards UE j, N_0 is the Gaussian noise power spectrum density. Also $G_{i,j}^{\text{Tx}}(t)$ and $G_{i,j}^{\text{Rx}}(t)$ are the transmit and receive antenna gain between BS i and UE j, respectively. To reflect the impact of the environment on channels, we define $\zeta_{i,j}(t)$ as the small-scale fading coefficient, $G_{i,j}^{\text{H}}(t)$ channel gain capturing the path-loss and large-scale shadowing effect. Eventually, $I_{i,j}^{(a)}(t)$ is the total intra- and inter-cell interference experienced by UE j communicating with BS i. Hence, the access capacity, $C_{i,j}^{(a)}(t)$, of the link between BS i and UE j reads as:

$$C_{i,j}^{(a)}(t) = (1 - \mu)B \cdot \log_2(1 + x_{i,j}(t)\text{SINR}_{i,j}^{(a)}(t)), \quad (2)$$

where $x_{i,j}$ is the binary UE association variable, which equals 1 when UE j is associated with BS i and 0 otherwise. Similarly, the backhaul capacity, $C_i^{(b)}(t)$, of the link between MAP i and the IAB donor reads as:

$$C_i^{(b)}(t) = \mu B \cdot \log_2(1 + z_i(t)\text{SINR}_{i,j}^{(b)}(t)), \quad (3)$$

where we define $z_i(t)$ as the binary backhaul link association variable, which indicates if a MAP is currently deployed or not. Here, the $\text{SINR}_i^{(b)}(t)$ experienced by the MAP i communicating with the IAB donor is given by:

$$\text{SINR}_i^{(b)}(t) = \frac{\zeta_{0,i}(t) P_{0,i}^{\text{Tx}} G_{0,i}^{\text{Tx}}(t) G_{0,i}^{\text{H}}(t) G_{0,i}^{\text{Rx}}(t)}{I_i^{(b)}(t) + \mu N_0 B}, \quad (4)$$

where, $I_i^{(b)}(t)$ denotes the intra-backhaul interference.

It is worth noting that in Eq. (2)-(3), the SINR and the channel capacity depend on path losses and interference influenced by various topological factors. Our system model considers ground-to-ground and air-to-ground mmWave path loss, which are affected by Line-of-Sight (LoS) conditions and the distance $d_{i,j}(t) = \|\ell_i(t) - \ell_j(t)\|$ between MAP i and UE j at time t. We omit full description here due to lack of space and refer readers to our previous work [14].

B. Effective Rate and Network Sum-rate

Let $D_j(t)$ define the traffic request of UE j at time t (in bps). Hence, $\Gamma_{i,j}(t) = \min(D_j(t), C_{i,j}^{(a)}(t))$ represents the effective data requirement on the access link between UE j and MAP i. Thus, if $\beta_{i,j}(t) \in [0, 1]$ is the fraction of MAP i backhaul capacity $C_i^{(b)}(t)$ allocated to UE j, the instantaneous effective rate $R_{i,j}(t)$ perceived by UE j from BS i reads as:

$$R_{i,j} = \begin{cases} \min(\Gamma_{i,j}(t), \beta_{i,j}(t) z_i(t) C_i^{(b)}(t)), \forall i \in \mathcal{S}, \\ \Gamma_{i,j}(t), \qquad\qquad\qquad\qquad\quad \text{if } i = 0. \end{cases} \quad (5)$$

Finally, we define the total network sum-rate $R(t)$ as:

$$R(t) = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{U}(t)} R_{i,j}(t) \quad (6)$$

III. PROBLEM FORMULATION

Our goal is to optimize the user experience in this dynamic networks with varying traffic demand, locations, and numbers of MAPs and UEs. We aim to optimize at the same time i) the number of deployed MAPs, ii) their backhaul allocation and iii) the dynamic placement of each MAP. To do so, we

formulate the Multi-MAP management problem to maximize the long-term sum rate as follows:

$$\max_{\boldsymbol{\Psi}(t)} \quad \lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[R(t)], \tag{$\mathcal{P}$}$$

$$\text{s.t.} \quad x_{i,j}(t), z_i(t) \in \{0,1\}, \qquad \forall i \in \mathcal{S}_0, j \in \mathcal{U}(t), \tag{$\mathcal{C}_1$}$$

$$\sum_{j \in \mathcal{U}(t)} x_{i,j}(t) \le K_i(t), \qquad \forall i \in \mathcal{S}_0(t), \tag{$\mathcal{C}_2$}$$

$$\sum_{i \in \mathcal{S}_0} x_{i,j}(t) \le 1, \qquad \forall j \in \mathcal{U}(t), \tag{$\mathcal{C}_3$}$$

$$M_s(t) = \sum_{i \in \mathcal{S}} z_i(t) \le M, \tag{$\mathcal{C}_4$}$$

$$\beta_{i,j}(t) \in [0,1], \qquad \forall i \in \mathcal{S}, j \in \mathcal{U}(t), \tag{$\mathcal{C}_5$}$$

$$\sum_{j \in \mathcal{U}(t)} \beta_{i,j}(t) \le 1, \qquad \forall i \in \mathcal{S}(t), \tag{$\mathcal{C}_6$}$$

$$\ell_i(t) \in \mathcal{L} \subset \mathbb{R}^3, \qquad \forall i \in \mathcal{S}(t), \tag{$\mathcal{C}_7$}$$

$$\|\ell_i(t+1) - \ell_i(t)\| \le \Delta\ell, \qquad \forall i \in \mathcal{S}(t), \tag{$\mathcal{C}_8$}$$

where $\boldsymbol{\Psi}(t) = \{z_i(t), \beta_{i,j}(t), \ell_i(t), \forall i, j\}$ are the optimization variables and the expectation in ($\mathcal{P}$) is taken *w.r.t.* the random processes, whose statistics are unknown. In ($\mathcal{P}$), constraint ($\mathcal{C}_2$) ensures that each BS i serves at most $K_i(t)$ UEs simultaneously. Constraint ($\mathcal{C}_3$) guarantees that each UE is associated to one BS at a time. Similarly, ($\mathcal{C}_4$) ensures that the IAB donor serves at most M active backhaul links simultaneously. Moreover, ($\mathcal{C}_5$)-($\mathcal{C}_6$) guarantees a positive backhaul allocation $\beta_{i,j}(t)$ for each UE i connected to MAP $i \in \mathcal{S}(t)$ and sum to at most one at each time t. Finally, regarding MAPs mobility, ($\mathcal{C}_7$)-($\mathcal{C}_8$) define a bounded region $\mathcal{L}$ of space where MAPs cannot move more than $\Delta\ell$ meters at a time. Problem ($\mathcal{P}$) is a non-convex combinatorial problem whose complexity increases with network size. In addition, there is an interdependence in optimization variables. Indeed, the required number of MAPs depends on UE topology, such as location and traffic demand distribution, which determines whether a dense or scattered deployment is necessary. For determining the optimal MAP locations, a centralized exhaustive search is not feasible due to interdependence between the number and locations of MAPs and the complexity of the network's interference profile. Following [18], given the user association $x_{i,j}(t)$, here specified by the max-SNR algorithm, the backhaul capacity allocation $\beta_{i,j}(t)$ can be obtained using convex optimization. Using max-SNR algorithm, the values of $x_{i,j}(t)$ are defined by the locations and number of MAPs. Therefore, in the remaining, we focus on finding $\{z_i(t), \ell_i(t)\}$. To solve this problem with limited complexity, we propose a two-level hierarchical optimization framework to optimize $\boldsymbol{\Psi}(t)$.

IV. PROPOSED SOLUTION

We propose a two-level framework to address problem ($\mathcal{P}$), wherein the high-level is responsible for jointly determining the number of MAPs and the low-level uses federated MADRL to position the MAPs under time-varying network constraints. As illustrated in Fig. 2, the high-level gathers network information $o^{(h)}(t)$ (1) and feedbacks of MAPs currently deployed to make the **MAP number decision** (2). During the training phase, the high-level determines the best target location $\ell_i^*(t)$
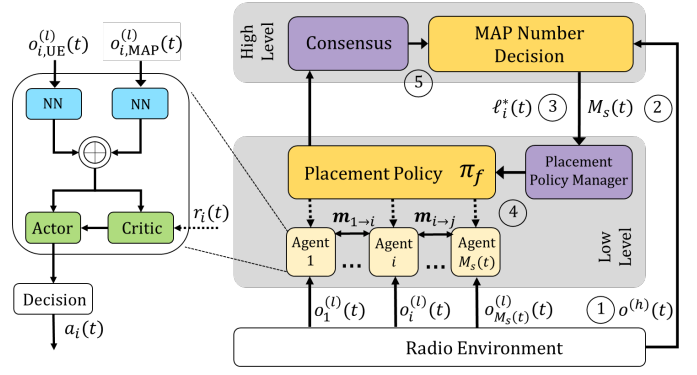

Fig. 2. Proposed hierarchical architecture for 3D operation of MAPs.

for each MAP i, which is used to compute MAPs agent training rewards (3). These locations are no longer transmitted afterwards, *i.e.* they only serve for training, since the agents have learned to determine them on their own. Then, the low-level **placement policy manager** loads the **placement policy** to the deployed MAPs and sends them to the network. For the training phase, each MAP federates its local model to a common placement policy π_f (4). Finally, each MAP determines its relative importance within the network based on its trade-off to achieve a **Consensus**. Thus, each MAP decides whether to repatriate or enable a new MAP for assistance in serving the UEs (5). This dynamic deployment of MAPs is iterative, scalable, and distributed.

A. High-Level - Decentralized Trade-off

In this section, we discuss the intuitions and motivations behind the proposed distributed **Trade-off algorithm**. First, the MAP number issue must be addressed simultaneously with topology and radio configuration conflict constraints. In fact, deciding the number of MAP to deploy depends on: i) the network **topology** *i.e.* UE's distribution; ii) the network nodes **configuration** *i.e.* UE's association; which are not considered in standard approaches like clustering. We propose to include both aspects in the calculation of the trade-off θ_i specific for each MAP i. Thus, the associated problem raises multiple challenges: the self-dependency of MAP locations and number, ensuring sufficient UE coverage, maintaining backhaul connectivity, minimizing the number of MAPs to limit operational complexity. To tackle these objectives, with low complexity and a long-term vision, we propose an iterative algorithm that considers both aspects in the trade-off calculation.

Trade-off computation. As described in Algorithm 1, each MAP i maintains locally a trade-off value $\theta_i(t)$ to decide if it needs a support from a new MAP or to be repatriated. Therefore, decision making is decentralized and the number of MAPs is determined by consensus. Then, each MAP computes its UEs inertia $\Phi_i(t)$, determined by the sum of squared distance of active UEs to the MAP, which captures the network topology surrounding the MAP. Additionally, the MAP i determines if its beams are currently overloaded with served UE or lower than a threshold number of beam $K_{i,\min}$, which captures the current network configuration. We

---

Algorithm 1: `Trade-off Algorithm`

Input: $\mathcal{S}$ set of MAPs; $\mathcal{U}$ set of UEs
K_i maximum beam per MAP

1 Enable $K(t)/\mathbb{E}_i[K_i]$ MAPs and start low-level MADRL algorithm [14]
2 **for** $t \in [0, T_n]$ **do**
3   **if** t modulo $\tau_n = 0$ **then** Init $\theta_i = \{i : 0\}\forall i \in \mathcal{S}$
4   **for** $i \in \mathcal{S}$ **do**
5   **if** $\overline{\theta_i} < 0$ **then** MAP i decides to repatriate
6   **if** $\overline{\theta_i} > 0$ **then**
7   MAP i activates MAP i^*; $z_{i^*} = 1$
8   Placement Policy Manager sends π_f to MAP i^*
9   **if** $M_s(t) < M$ **then**
10   Update backhaul capacity allocation $\beta_{i,j}$
11   Update UE associations $x_{i,j}$
12   **for** $i \in \mathcal{S}$ **do**
13   Update θ_i through *local monitoring*

TABLE I
APPROACHES COMPARISON

Benchmark	**CODEBOOK**	**Curriculum**	**Federated**
Context	Context-aware (Specialized)	Context-free (Generalized)	
Complexity ($\mathcal{O}_c$)	$\frac{M(M+1)}{2}$	M	1
Policy	$\{\pi_{k,i}\},$ $\forall k \in \{0,...,i\}, \forall i \in \mathcal{S}$	$\{\pi_i\}, \forall i \in \mathcal{S}$	π_f

find here the interdependence of the optimization variables since the management of the beams is ensured by deciding $x_{i,j}$ and z_i. When the inertia is high or when the station is overloaded, θ_i is increased for each aspect and decreased in the opposite cases. Both metrics are acquired via *local monitoring* of UEs within its coverage range and averaged to guarantee a long-term vision. Notice that each MAP must set up and execute the low-level hierarchy related to the current network configuration, which implies an operational cost. In fact, the MADRL framework guarantees only a limited generalization ability as it must have constant size input information. In consequence, we propose a new approach based on a federated learning mechanism to unify every MAP model into a single model, easy to maintain.

B. Low Level - Cooperative Placement

To solve the dynamic MAP placement problem, we propose to model each MAP as an autonomous agent that have to co-operate to serve a dynamic 5G network. This approach comes with new challenges: follow and distribute UEs demand; schedule their path over time; collect and process surrounding information perception by their own. For this purpose, we propose a Multi-Agent Deep Reinforcement Learning (MADRL) algorithm as the low-level of our hierarchical architecture. Thus, to efficiently solve the MADRL problem, we proposed in [14] a double-attention actor-critic architecture. This model achieves a distributed cooperation without any prior information and without retraining procedures for time-varying scenarios. This cooperation is accomplished by learning, ex-changing, and interpreting messages $\mathrm{m_{i,j}}$ between agents. The proposed solution solves multiple challenges: i) model-free property for the incoming radio environment; ii) agent state observations efficient representation; iii) network scalability; iv) distributed cooperation. In this approach, each MAP is

modelled as an agent, which continuously learns to make autonomous decisions based on partial observations $o_{i,\mathrm{UE}}^{(l)}(t)$ from grounded UEs $\mathcal{N}_{i,\mathrm{UE}}$ and the messages $o_{i,\mathrm{MAP}}^{(l)}(t)$ received from other deployed MAPs $\mathcal{N}_{i,\mathrm{MAP}}$. We formalize this decision process as a Markov Decision Process (MDP). Each time slot, agents receive UEs location $\ell_j(t)$ and other MAPs location $\ell_i(t)$ to build a context representation $o_i^{(l)}(t)$. Based on this observation, agents select actions from a predefined set $\mathcal{A}$ = {forward, backward, up, down, left, right, hover}, corresponding to movement of the associated MAPs along the selected direction with a fixed step size $\Delta \ell$. Agents then transition to new observations $o_i^{(l)}(t+1)$ and receive rewards according to the following multi-objective reward function:

$$r_i(t) = (\delta_i(t) - 1)d_i(t) + \delta_i(t)(C_i^{(b)}(t) - d_0). \quad (7)$$

Here, $\delta_i(t) = \mathbb{1}(d_i(t) \le d_0)$, where d_0 is a reference distance and $d_i(t) = \|\ell_i(t) - \ell_i^*(t)\|$ is the distance of MAP i to its optimal location $\ell_i^*(t)$. Since this location is not known a priory, we approximate it during the training phase with the location of the nearest assigned centroid obtained by clustering UEs using *e.g.* `Kmeans` algorithm. As demonstrated in our previous work [18], this multi-objective reward pushes agents to maximize user coverage and backhaul capacity at the same time. Each agent then learns a policy $\pi_i(t)$ that maxi-mizes the expected sum of perceived (γ-discounted) rewards $\mathbb{E}_\pi[\sum_{\tau=t}^{T_l} \gamma^{\tau-t} r_i(\tau)]$ over a time horizon T_l, where $\gamma \in [0, 1)$. However, as the dynamic of the network evolves, new training mechanisms are required to maintain network performance. Frequent training processes are prohibitive, induce latency, and a signalling overhead, which are detrimental to network operation efficiency. Therefore, new approaches are required to learn MAP placement policies, which are i) context-free *i.e.* independent of the number of deployed MAPs $M_s(t)$, ii) scalable to cope with size-varying number $K(t)$ and position of UEs, iii) generalizable to different network deployment, which is a current and fundamental topic in MADRL, iv) and with limited operational complexity ($\mathcal{O}_c$).

Trivial approach via a codebook of policies. With the varying number of MAPs, the first trivial approach, which will serve as **BASELINE** is to maintain a representative set of scenario-specific policies to form a **CODEBOOK**. This approach devises specialized models for every combination of the number of deployed MAPs: $\{\pi_{k,i}\}, \forall k \in \{0, ..., i\}, \forall i \in \mathcal{S}(t)$. Then, depending on the scenario, the **placement policy manager** selects the appropriate policy within the codebook to deploy. Obviously, this approach is complex and context-aware as it requires identification of the facing scenario, and maintaining $\mathcal{O}_c = \frac{M(M+1)}{2}$ different policies, where M is the maximal number of MAPs. In addition, this approach may fail to generalize to unseen scenarios, which may not be captured by the codebook. Thus we propose to reduce the number of policies to maintain and at the same time increase their generalization ability.

Share to conquer: a curriculum approach. In the context of a varying number of agents, we propose a curriculum MADRL (referred to **C-MADRL**) training approach. In contrast to the previous approach, each agent maintains its own model

Channel Parameters	IAB donor	MAP
Carrier Frequency f_c	2 GHz	28 GHz
Antenna Aperture Angle	180	90
Shadowing Variance σ_l^2	3 dB	12 dB
Antenna Gain	17 dBi	Directive [14]
Beam Forming	$K_0 = \infty$	$K_i = 10$
Thermal Noise N_0	-174 dBm/ Hz	
Small-Scale fading (m-Nakagami)	$m = 3$	
d_0	10	
Bandwidth partition	0.75	
$\Delta\ell$	5	
$\{\tau_f, \alpha_f\}$	$\{5000, 0.5\}$	
System Bandwidth B	500 MHz	
Learning rate	10^{-4}	
γ	0.6	
$\{N_{i,\text{UE}}, N_{i,\text{MAP}}\}$	$\{15, 5\}$	
$D_i(t)$ (k-Poisson distribution)	$k = 1$ Gbps	

Fig. 3. Hierarchical Learning convergence of low-level policies.

through all possible configurations, which reduces the operational complexity to $\mathcal{O}_c = M$. During the training procedure, we randomly sample a scenario with a random number of deployed MAPs. Then, the deployed MAPs cooperatively learn their respective policies, which we maintain across a different sampling of scenarios. This method is context-free and allows each MAP agent to generalize to different scenarios with a different number of deployed MAPs thus fostering cooperation with size-varying teammates.

A transferable policy via federated mechanism. Here, we propose a Federated MADRL (referred to **F-MADRL**) mechanism. The goal is to share the knowledge of placement and cooperation into a single policy π_f that can be propagated to new any agent, no matter their number no matter which agent is enabled, which reduces the operational complexity to $\mathcal{O}_c = 1$. This approach brings the MAP placement problem to a new dimension where the issue is no longer to determine the architecture of the models but the processing of observations and cooperation. Then, contrary to **C-MADRL**, where each agent can be distinguished fundamentally by its model, this approach introduces the new challenge of distinguishing agents based solely on observations. To achieve this, during the training phase [14], the federated mechanism retrieves the weights of all the agent models $w_i(t)$ to average them and updates the agent models with a proportion rate α_f every τ_f: $w(t) = \alpha_f \times w(t) + \frac{(1-\alpha_f)}{M_s(t)} \times \sum_i w_i(t)$. In the proposed solution, parameters α_f and τ_f ensures the model stability while generalizing, avoiding lack of convergence.

The federation of models during the training guarantee that the resulting policy is transferable irrespective of the scenario and the number of deployed MAPs.

To assess the performance of the aforementioned approaches, we introduce a new metric, termed the operational efficiency, which we define as $\eta(t) = \frac{R(t)}{\mathcal{O}_c}$, where $\mathcal{O}_c$ is the operational complexity defined by the number of different policies to maintain for achieving $R(t)$ (see Table I).

V. NUMERICAL RESULTS

In this section, we evaluate the performance of our two-level hierarchical framework in a dynamic 5G network. To do so, following our previous work [14], we train policies using actor-critic framework with proximal policy optimization (PPO) [19]. We refer readers to [14] for detailed description.
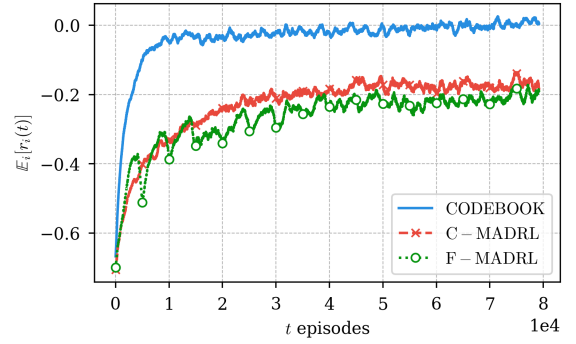
For the **CODEBOOK** construction, we train a set of model for scenarios with $\{2, 3, 4\}$ MAPs. This approach may be assimilated to the standard *state-of-the-art* approach with specialized models that do not take into account a variable number of MAP. For the exploitation phase, when $M_s(t) > 4$, a random model is sampled from the $M_s(t) = 4$ codebook. For the **C-MADRL** and the **F-MADRL** training, we randomly deploy from 2 to 5 MAPs on a random locations sampled in a 200 m by 200 m area, where $K(t) = 25$ UEs are deployed. Table II summarizes simulation parameters.

Federation Policy Convergence. To begin, we assess convergence performances of proposed benchmarks. Fig. 3 shows the rolling averaged reward over a 500-sized window and over all agents. Under the constraints of a single policy, the **F-MADRL** solution is able to acquire the capacity to cooperate within a single policy as it have the same convergence than the **C-MADRL** and **CODEBOOK** approaches. Though there are drops in reward due to the federation mechanism, it stabilizes during training, confirming the acquisition of cooperation capacity in one single policy. However, due to the generalization capability provided by the federation, the observed reward is lower compared to the specialized **CODEBOOK** approach, which is specialized for a every scenario.

Federation for Generalization. We examine every MADRL generalization ability in the dynamic 5G network. For 200 configurations that last $T_l = 100$ iterations, we deploy now $K(t) = 60$ UEs, which does not correspond to any training scenario and $M_s = K(t)/K_i$ MAPs at $t = 0$. UEs now follow a random way-point centroid mobility at 0.8m/s with a blockage probability of 0.5 that leads to a variable total number of connected UE and MAPs between each episodes. As every model has not been trained with specific mobility model, it is able to support multiple type of mobility. Fig. 4 compares the averaged sum-rate achieved $\mathbb{E}[R(t)]$ for different network scenarios. Here, the **CODEBOOK** approach suffers from a drop of performance in unseen scenarios, while the **F-MADRL** continuously increases and scales with the network with a 31% improvement with $M_s(t) = 6$, while the **C-MADRL** stabilizes its performances with $M_s(t) \geq 4$. Most importantly, we examine the operational efficiency $\mathbb{E}[\eta(t)]$ and we observe that the loss of performance in the training process to increase the generalization capacity of the **F-MADRL** single policy is largely compensated by its cost of exploitation.
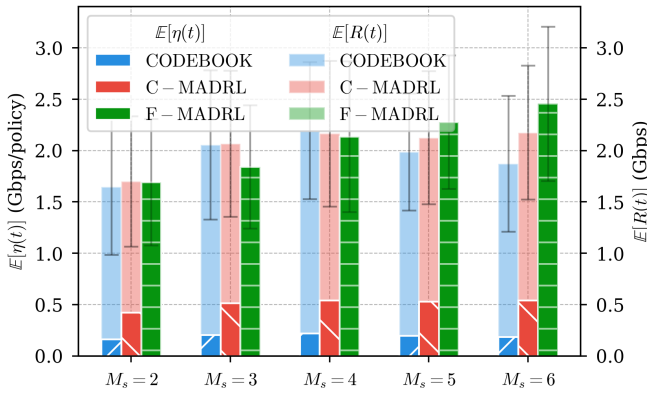
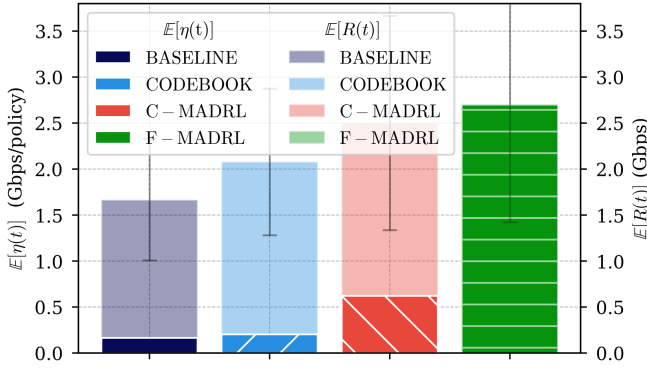Fig. 4. Performance comparison between proposed approaches.

Fig. 5. Performance comparison compared to the baseline.

Performance Comparison with Fixed Number of MAPs.
Here, we use the proposed `trade-off algorithm` to dynamically manage the MAP number within the same episode every $t_n = 10$. We set inertia thresholds to $\Phi_{i,\max} = 6 \times 10^3$, $\Phi_{i,\min} = 0$ and MAP i loads thresholds $K_{i,\min} = 2$, $K_i(t) = 10$. For each threshold, the trade off of each MAP $\theta_i(t)$ is increased or decreased by 1. Fig. 5 compares $\mathbb{E}_i[R(t)]$ and $\mathbb{E}_i[\eta(t)]$ for all network configurations encountered and MAP i. Thus, compared to the state-of-art **BASELINE**, which consider a codebook with a **fixed** number of MAP deployed within an episode, our two-level hierarchical framework achieves an increase of 62% of the averaged sum-rate while demonstrating its operational efficiency. Moreover, the introduction of a dynamic number of MAP for the **CODEBOOK** approach results in a 24% increase of $\mathbb{E}[R(t)]$, which confirms the need for MAP number adjustment in Multi-MAP networks to meet 5G ambitions. This sum-rate increase can be explained by the better management of UE mobility and interference. As a result, our solution is able to guarantee a better performance even with a high number and density of UEs.

VI. CONCLUSION

This study proposes a scalable and distributed solution for determining the optimal placement and number of MAPs in a dynamic 5G network with IAB constraint. The solution utilizes a two-layer hierarchical approach where MAPs decide on their number and optimize backhaul connectivity while autonomously reconfiguring the network. Numerical evaluations show up to 62% network sum-rate increase and improved operation efficiency compared to a state-of-the-art baseline. The proposed solution removes the constraint for a fixed number of deployed MAPs, paving the way for more realistic multi-agent systems with a varying number of agents.

REFERENCES

[1] M. Maman, E. Catte, M. Sana, M. Girmay, V. Maglogiannis, *et al.*, "Coverage Extension as a Service for Flexible 6G Networks Infrastructure," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 1329–1334, 2022.

[2] M. Peer, V. A. Bohara, *et al.*, "User Mobility-Aware Time Stamp for UAV-BS Placement," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, 2021.

[3] R. Ghanavi, E. Kalantari, M. Sabbaghian, *et al.*, "Efficient 3D Aerial Base Station Placement Considering Users Mobility by Reinforcement Learning," *CoRR*, vol. abs/1801.07472, 2018.

[4] L. Wang, H. Zhang, S. Guo, and D. Yuan, "3D UAV Deployment in Multi-UAV Networks With Statistical User Position Information," *IEEE Communications Letters*, vol. 26, no. 6, pp. 1363–1367, 2022.

[5] S. Sharafeddine and R. Islambouli, "On-Demand Deployment of Multiple Aerial Base Stations for Traffic Offloading and Network Recovery," *CoRR*, vol. abs/1807.02009, 2018.

[6] J. Sabzehali, V. K. Shah, Q. Fan, B. Choudhury, L. Liu, and J. H. Reed, "Optimizing Number, Placement, and Backhaul Connectivity of Multi-UAV Networks," *CoRR*, vol. abs/2111.05457, 2021.

[7] B. Zhang, J. Song, *et al.*, "Genetic algorithm enabled particle swarm optimization for aerial base station deployment," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pp. 1–7, 2021.

[8] 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, 2017.

[9] J. Qin, Z. Wei, C. Qiu, and Z. Feng, "Edge-prior placement algorithm for uav-mounted base stations," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2019.

[10] Q. Wu, Y. Zeng, and R. Zhang, "Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, 2018.

[11] N. Zhao, Z. Liu, and Y. Cheng, "Multi-Agent Deep Reinforcement Learning for Trajectory Design and Power Allocation in Multi-UAV Networks," *IEEE Access*, vol. 8, pp. 139670–139679, 2020.

[12] Z. Qin, Z. Liu, G. Han, C. Lin, L. Guo, and L. Xie, "Distributed UAV-BSs Trajectory Optimization for User-Level Fair Communication Service With Multi-Agent Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, 2021.

[13] S. Zhou, Y. Cheng, X. Lei, *et al.*, "Resource Allocation in UAV-assisted Networks: A Clustering-Aided Reinforcement Learning Approach," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2022.

[14] E. Catté, M. Sana, and M. Maman, "Dual-Attention Deep Reinforcement Learning for Multi-MAP 3D Trajectory Optimization in Dynamic 5G Networks," *arXiv preprint arXiv:2303.05233*, 2023.

[15] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Meta-Reinforcement Learning for Trajectory Design in Wireless UAV Networks," *CoRR*, vol. abs/2005.12394, 2020.

[16] N. Iradukunda, Q.-V. Pham, M. Zeng, H.-C. Kim, and W.-J. Hwang, "UAV-Enabled Wireless Backhaul Networks Using Non-Orthogonal Multiple Access," *IEEE Access*, vol. 9, pp. 36689–36698, 2021.

[17] Y. Dai, Y. Guo, and J. Hao, "UAV Placement and Resource Allocation for Multi-hop UAV Assisted Backhaul System," in *proc. IEEE Conference on Computer Communications Workshops*, pp. 1–6, 2021.

[18] M. Sana and B. Miscopein, "Learning Hierarchical Resource Allocation and Multi-agent Coordination of 5G mobile IAB Nodes," *arXiv preprint arXiv:2302.07573*, 2023.

[19] M. Sana, N. di Pietro, and E. C. Strinati, "Transferable and Distributed User Association Policies for 5G and Beyond Networks," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 966–971, 2021.