Distributed DRL with Multiple Learners for AP Clustering in Large-scale Cell-Free Deployment

Akio Ikami [1], Yu Tsukamoto [2], Takahide Murakami [3], Hiroyuki Shinbo [4], Yoshiaki Amano [5]
KDDI Research, Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, Japan.
([1] ak-ikami, [2] yu-tsukamoto,[3] tk-murakami, [4] hi-shinbo, [5] yo-amano) @kddi.com

***Abstract*— This paper proposes a distributed deep reinforcement learning (DRL) method with multiple learners for AP clustering in large-scale Cell-Free massive MIMO (CF-mMIMO). In the deployment of large-scale CF-mMIMO with many user equipments (UEs) and access points (APs), it is necessary to perform AP clustering according to the demand and movements of each UE in a lightweight manner and with high inference accuracy. However, existing DRL-based methods have struggled to learn diverse and site-specific radio environments and provide high inference accuracy with small amounts of data and small neural network (NN) models for lightweight. To address this problem, the proposed method classifies learners for each radio environment with the Reference Signal Received Power (RSRP) between surrounding APs of the UE to perform learning and inference with these multiple learners. Furthermore, the proposed method dynamically adjusts the association between UE and learners based on the fluctuation in RSRP due to the UE's movements, thereby ensuring sufficient agility for user mobility. This dynamic association of UEs and learners for each radio environment enables efficient learning and improved inference accuracy by focusing on UEs in similar radio environments, even with small amounts of data and small NN models. Simulation evaluations based on actual urban structures demonstrated that the proposed method realizes AP clustering with higher inference accuracy than existing methods, even with small amounts of learning data and small NN models.**

***Index Terms*—6G, RAN management, deep reinforcement learning, Cell-free massive MIMO.**

I. INTRODUCTION

Numerous organizations have identified robotics as a key use case for the sixth-generation mobile communication system (6G), anticipated for commercialization around 2030 [1], [2]. Mission-critical applications involving remote operation, robots replacing human labor, and utilization of Urban Air Mobility are key examples. In these 6G use cases, ensuring safety is one of the most critical requirements, necessitating constant monitoring and control of robots via 6G, regardless of location. This implies a strong demand for consistently high radio quality anytime and anywhere. However, the fifth-generation systems have issues at the cell edge due to increased path loss and inter-cell interference, leading to degradation of radio quality.

Cell-free massive MIMO (CF-mMIMO) is a promising technology that addresses the cell edge problem [3]. It utilizes access points (APs) around user equipment (UE) and a central processing unit (CPU) for signal processing, effectively suppressing inter-cell interference. Considering the urban and large-scale deployment of CF-mMIMO, the computational load for signal processing is extremely high. To reduce this computational load, AP clustering per user has been proposed [4]. Here, an AP cluster is a set of APs that transmit and receive radio signals for each UE. As the radio environment and required radio quality vary for each UE, a dynamic selection of APs for the AP cluster is necessary. Optimization of AP clustering using a mathematical approach has been proposed to maximize spectrum efficiency or guarantee throughput for each UE [5], [6], [7]. However, the optimization for AP clustering is a non-linear, non-convex problem due to the complexity of inter-UE interference [6]. So, this optimization approach results in a computational complexity problem for AP cluster decisions in actual environments where UEs move.

On the other hand, the application of deep reinforcement learning (DRL) to the AP clustering problem is being discussed [8], [9], [10], [11]. In DRL, an agent uses a neural network (NN) to understand the system model and finds the optimal control through trial and error [12]. In existing methods [8], [9], [10], the agent's actions are defined as arbitrarily selecting a combination of APs for each UE. The size of the action space, which is the number of possible actions, increases exponentially with the number of UEs and APs. In large-scale environments with many APs and UEs, a large action space, i.e., a large NN size, increases the complexity of learning and decision-making time for AP clustering. In [11], the authors design distributed DRL for AP clustering, which reduces decision-making time by distributing the actors. However, in large-scale environments, the inference accuracy for AP clustering is degraded because it is difficult to learn the diverse and site-specific radio environment for each UE sufficiently with a realistic computational load. This degradation becomes more pronounced in lightweight online learning with a small-sized NN and less data, posing a challenge for the large-scale deployment of CF-mMIMO.

To address this problem, this paper proposes a distributed DRL method for radio access network (RAN) management with multiple learners for each radio environment. In the proposed method, the radio environment is classified into multiple categories based on the measurements of the UEs, and classified learners are associated with each radio environment. Each learner learns the UE's experience in each radio environment, and each model tied to the radio environment infers for UEs that have similar UE measurements used for training. Since the proposed method uses UEs in similar radio environments for training, it can learn efficiently with less data and small NN models and ensure agility when changes in the radio environment occur. In addition, inference is performed on UEs in similar radio environments according to UE measurements, which improves inference accuracy. For the classification of the radio environment, we use the Reference Signal Received Power (RSRP) between surrounding APs of the UE, as defined in 3GPP and O-RAN. The proposed method is highly feasible

979-8-3503-8741-4/24/$31.00 ©2024 IEEE

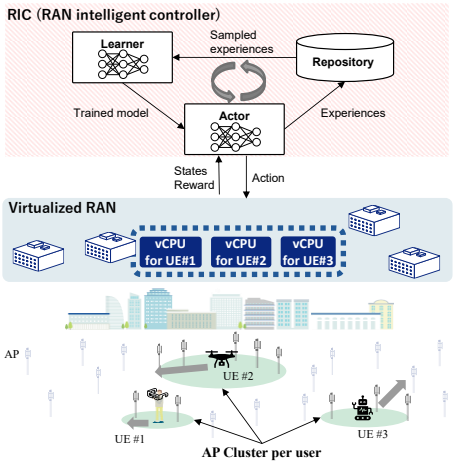

Fig. 1. RAN management architecture for assuring appropriate radio quality everywhere with CF-mMIMO.

since RSRPs can be obtained anytime in 5G compatible UEs and O-RAN compatible RAN Intelligent Controller (RIC). To describe the proposed method in detail, Section II explains the system model of this paper and the challenges when using DRL for AP clustering. Then, Section III describes the architecture and design of the proposed distributed DRL with multiple learners. In Section IV, simulation evaluations in site-specific environments based on actual urban structures show that the proposed method realizes AP clustering with higher inference accuracy than existing methods, even with less learning data and smaller NN models.

II. System Model

A. Architecture

Figure 1 introduces RAN management architecture with virtualized RAN to ensure appropriate radio quality everywhere using CF-mMIMO. The concept of this architecture is to generate and manage a logical network for each UE on a physical infrastructure based on network virtualization, relying on virtualized CPU (vCPU) and AP clustering. The vCPU processes radio signals from APs within the AP cluster to the UE. The RIC, responsible for controlling the logical network, manages radio quality by controlling the AP cluster according to each UE's mobility and requirements. The RIC manages to optimize AP selection for AP clustering that reduces the computational load involved in signal processing while satisfying the UE's throughput requirements. However, selecting the APs to form this AP cluster is a non-linear, non-convex problem due to the complexity of inter-UE interference [6], making it difficult to determine the optimal solution. Therefore, we consider the application of AI/ML, especially DRL, to find sufficiently near-optimal solutions within realistic computation time. AI/ML adaptation to RAN management has attracted considerable attention, and its application is being actively worked on in O-RAN, where RIC standardization is being discussed [13]. In this paper, the UE antenna is assumed to be single.

To select the appropriate APs for AP clustering when many UEs are moving and the radio environment is changing moment by moment, model updates based on online learning from UE measurements are necessary. The upper side of Fig. 1 shows the management of AP clustering by online learning using reinforcement learning. In the DRL functions within the RIC, the actor observes the state from the environmental information measured by the UE and AP and takes action. As an evaluation of the action, the actor receives a reward from the environment. The state, next state, action, and reward are stored in the repository as experiences. Concurrently, the learner learns from experiences sampled from the repository, and the actor updates the model with the learned model from the learner at regular intervals.

B. Mathematical formulation

In this section, we formulate the system model for AP clustering in CF-mMIMO. We consider K single-antenna UEs in an area in which L APs are deployed. The AP index, which belongs to an AP cluster for UE k, $\boldsymbol{D}_k$, is defined as the following L -dimensional square matrix,

$$\boldsymbol{D}_k = \begin{bmatrix} D_{k1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_{kL} \end{bmatrix}, \tag{1}$$

where D_{kl} is defined as

$$D_{kl} = \begin{cases} 1 & \text{if AP } l \text{ serves UE } k, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Let M_k, which is the set of APs when $D_{kl} = 1$, be the AP cluster for UE k. The SINR of the uplink for UE k is defined as

$$\mathrm{SINR}_k^{\mathrm{UL}} = \frac{p_k \left| \boldsymbol{v}_k^H \boldsymbol{D}_k \widehat{\boldsymbol{h}}_k \right|^2}{\sum_{i=1, i\neq k}^{K} p_i \left| \boldsymbol{v}_k^H \boldsymbol{D}_k \widehat{\boldsymbol{h}}_i \right|^2 + \boldsymbol{v}_k^H \boldsymbol{Z}_k \boldsymbol{v}_k}, \tag{3}$$

where $\boldsymbol{Z}_k = \boldsymbol{D}_k (\sum_{i=1}^{K} p_i \boldsymbol{C}_i + \sigma^2 \boldsymbol{I}_L) \boldsymbol{D}_k$, p_i is the power of the uplink signal, and $\widehat{\boldsymbol{h}}_i$ is the estimated channel coefficient. The channel coefficients are estimated with the minimum mean square error (MMSE) based on the pilot assignment method [4]. $\boldsymbol{C}_i$ is the matrix of the channel estimation error for UE i, which is obtained from the difference between the spatial channel correlation matrix estimated with the MMSE and a real one. σ^2 is the power of thermal noise. $\boldsymbol{I}_L$ is the L-dimensional identity matrix. The combining vector $\boldsymbol{v}_k$ is given as follows:

$$\boldsymbol{v}_k = p_k \left(\sum_{i \in \mathcal{P}_k} p_i \boldsymbol{D}_k \widehat{\boldsymbol{h}}_i \widehat{\boldsymbol{h}}_i^H \boldsymbol{D}_k + \boldsymbol{Z}_k \right)^{\dagger} \boldsymbol{D}_k \boldsymbol{h}_k \tag{4}$$

where, $\mathcal{P}_k$ is the set of UEs where the AP cluster for the UE is formed with at least one AP as used in the AP cluster for UE k expressed as $\mathcal{P}_k = \{i : \boldsymbol{D}_k \boldsymbol{D}_i \neq \boldsymbol{O}_L\}$ where $\boldsymbol{O}_L$ is the L-dimensional zero matrix. The uplink throughput g_k for UE k is calculated with SINR as

$$g_k = W_{\mathrm{RF}} \log_2 \left(1 + \mathrm{SINR}_k^{\mathrm{UL}}\right) \tag{5}$$

where, W_{RF} is the total bandwidth of the wireless link.

The total computational load involved in the signal processing required for vCPU is defined as the computational load C^{comp}, which is given by the following equation [4]

$$C^{\mathrm{comp}} = \sum_{k \in K} \left(C_k^{\mathrm{est}} + C_k^{\mathrm{weight}} \right), \tag{6}$$

where C_k^{est} is the computational load required for the channel estimation of the UE k, C_k^{weight} is the computational load required for the weight vector calculation of UE k, which can be expressed as follows

$$C_k^{\mathrm{est}} = \left(N\tau_p + N^2\right) |M_k| |\mathcal{P}_k|, \tag{7}$$

$$C_k^{\text{weight}} = \frac{(N|M_k|)^2 + N|M_k|}{2}|\mathcal{P}_k| + (N|M_k|)^2 + \frac{(N|M_k|)^3 - N|M_k|}{3} \quad (8)$$

Here, N is the number of antennas deployed in the AP, and τ_p is the number of pilot sequences. The computational load for signal processing is proportional to the cube of $|M_k|$, which is the number of APs belonging to the AP cluster for UE k.

C. Problem statement and related works

From the perspective of MIMO diversity, the radio quality of the UE improves as the size of the AP cluster increases. However, as shown in (8), the computational load for signal processing increases with the cube of the number of APs contained in AP clusters, requiring the minimum necessary AP clustering according to user demand. Since the radio environment varies depending on the UE's location, changes in the AP cluster that follow user movement are needed. The application of DRL to the AP clustering problem is being considered [8], [9], [10]. In these methods, the agent's actions are defined as arbitrarily selecting a combination of APs for each UE. The size of the action space, which is the number of possible actions, increases exponentially with the number of UEs and APs. In large-scale environments with many APs and UEs, a large action space, i.e., a large NN size, increases the complexity of learning and decision-making time when selecting APs for AP clustering.

To address this problem, the method in [11] ensures scalability by defining an actor for each UE, thereby making the size of the actor NN independent of the number of UEs in the environment. This reduces the computational load for learning and inference. However, while the method works in an ideal propagation environment with only line-of-sight (LOS), the actual radio environments are complex, with many buildings and site-specific propagation characteristics. In large-scale environments with diverse and site-specific radio environments, it is challenging to sufficiently learn the site-specific radio environment with a realistic computational load, leading to a decrease in inference accuracy. Online learning in RAN management for CF-mMIMO, as shown in Fig. 1, requires agility and real-time inference capabilities, necessitating reduced computational load for AP clustering management with a small-sized NN and less training data. Therefore, a method that can learn diverse radio environments with less training data and a small NN while ensuring high inference accuracy is needed.

III. Proposed Method

A. Proposed distributed DRL with multiple learners

1) Architecture

To improve inference accuracy in large-scale and diverse radio environments, this paper proposes a distributed DRL method for RAN management with multiple learners for each radio environment. Figure 2 shows the architecture of the proposed method. There are multiple learners classified for each radio environment and actors defined for each UE. The learners and the actors for each UE are tied according to the radio environment based on the measurements of the UEs. The

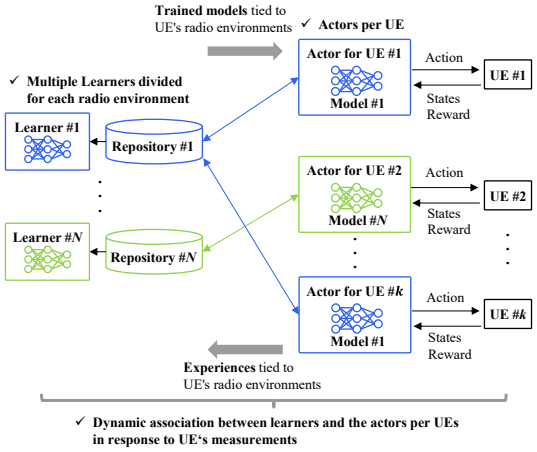

Fig. 2. Diagram of proposed distributed DRL with multiple learners.

experiences of UEs associated with each radio environment are input into the corresponding learner, and each learner executes the learning process individually. The learning model for each radio environment is distributed to the actor defined for each UE, and inference is performed according to the State. Since the UE's radio environment changes according to the UE's movement, we associate dynamically the ties between the UE and the learner with the UE's measurement information. This online learning process based on the UE's measurement information is repeated, enabling RAN control according to the constantly changing radio environment. Since the proposed method uses UEs in similar radio environments for training, it can learn efficiently with less data and small NN models. In addition, inference accuracy is improved because inference is performed using models learned from UEs in similar radio environments.

2) Design of multiple learners

This paper uses the RSRP, a measurement value defined by 3GPP, to classify radio environments. This measurement can be obtained from 5G compatible UEs, and the RIC can obtain it from O-RAN-compatible distributed units in RAN. Let $P_{l,k}$ be the value of the RSRP between UE k and AP l, and x_{ini} be the initial value of the number of APs associated with the AP cluster. The initial AP clusters are associated with the APs in order of increasing $P_{l,k}$. For the classification, we assume that the RSRP value follows a normal distribution when a large number of UEs are scattered over an area. We classify UEs into four groups $\boldsymbol{\mathcal{X}}_1 \ldots \boldsymbol{\mathcal{X}}_4$ and split and tie the learners to each of these groups. We calculate $P_k^s = \sum_{l\in(D_{kl}=1)} P_{l,k}$ and assume that P_k^s follows a normal distribution. Here these groups are defined based on the standard deviation σ_s as $\boldsymbol{\mathcal{X}}_1 = \{k| -\sigma_s > P_k^s\}$, $\boldsymbol{\mathcal{X}}_2 = \{k \mid -\sigma_s \le P_k^s \le 0\}$, $\mathcal{X}_3 = \{k \mid 0 \le P_k^s \le \sigma_s\}$, $\boldsymbol{\mathcal{X}}_4 = \{k \mid \sigma_s \le P_k^s\}$. By classifying the radio environment in this ratio-based way, we can balance the classification of diverse radio environments and the securing of training data. Although there is room to consider the number of divisions of the learner in detail in terms of the acquired amount of training data and inference accuracy, this is not covered in this paper, and we will leave it to future work.

B. Design for DRL for decision of AP clustering

1) State

For decision of AP clustering, we define the state for UE k as $S_k = \left[|M_k|^{\text{pre}}, \tilde{g}_k, R_k, R_k^{\text{pre}}, j_k\right]$. The previous AP cluster

size $|M_k|^{\text{pre}}$ is needed to determine the difference in the AP cluster size from the previous time step. The throughput requirement $\tilde{g}_k$ is needed to ascertain the required radio quality of the UE. $R_k = [r_{k,1}, r_{k,2}, \dots, r_{k,b}, \dots, r_{k,B}]$, where $r_{k,b} = P_{b,k} / \sum_{i=1, i\neq k}^{K} P_{b,\mathrm{i}}$ and it denotes the SNR from the b-th highest AP using RSRPs. R_k is an array of SNRs from APs arranged in descending order up to the B-th. R_k^{pre} is R_k at the previous time step. R_k^{pre} helps to learn the change in the channel state due to UE mobility. To consider the impact of other UEs around UE k, we employ j_k as the number of overlapping APs in the AP cluster of the UE k and other UEs. j_k is represented as $j_k = \sum_{i=1, i\neq k}^{K} |\boldsymbol{D}_k \boldsymbol{D}_i|$.

2) Reward

We aim to learn the minimum AP cluster size for each UE that meets the throughput requirements, depending on the radio environment. We define the reward r_k, which consists of two factors: throughput satisfaction and the AP cluster

$$r_k = q_k m_k \tag{9}$$

where, q_k and m_k are defined as

$$q_k = \begin{cases} 1 & g_k \geq \tilde{g}_k, \\ 0 & \text{otherwise.} \end{cases}, \; m_k = 1 - \left(\frac{|M_k|}{O}\right)^3.$$

The term q_k represents throughput satisfaction, where $\tilde{g}_k$ is the preset throughput requirement for UE k. If the throughput g_k does not meet the throughput requirement $\tilde{g}_k$ the reward becomes 0. The term m_k represents the factor in the AP cluster size. The computational load for signal processing is proportional to the cube of the AP cluster size $|M_k|$, as per equation (8). The term m_k decreases in proportion to the cube of the AP cluster size normalized by O. The reward is high when the throughput requirements are met with the minimum AP cluster size for UE k. The parameter O is a value larger than the maximum number of $|M_k|$, and the appropriate value of O changes depending on the number of pilot sequences. In this paper, we do not discuss this parameter and leave it for future work. In the simulations, we use empirically obtained values.

3) Action

We adopt an action design that specifies the increment/decrement in AP cluster size from the previous time step to increase or decrease the number of APs belonging to the AP cluster, as proposed in [11]. In an environment with densely distributed APs, the probability of rapid changes in radio quality due to large-scale fading is low, as many APs cover the UE. Therefore, it is sufficient for the actor to specify the difference in AP cluster size from the previous time step. The action for UE k is defined as $a_k = \delta_k \in \{-s, -s+1, \dots, 0, \dots, s-1, s\}$, where δ_k is the increment/decrement in the AP cluster size for UE k from the previous time step. Here, s represents the variation range by which the AP cluster size can increase or decrease in one time step. The AP cluster size is determined as $|M_k|$, where $|M_k|^{\text{pre}}$ represents the AP cluster size at the previous time step. The size of the action space is $2s + 1$.

IV. Performance Evaluation

Table I shows the computer simulation conditions and evaluation environment for urban cell-free deployment. To simulate a site-specific radio environment, we use a 1km^2 urban structure around Shibuya Station in Tokyo and employ path loss data based on ray tracing. To evaluate the proposed method, we compare it with the following three methods:

TABLE I: Simulation parameters

RAN environment parameters	
Simulation area	1km×1km at Shibuya in Tokyo
Number of deployed APs and UE	400, 100
Number of antennas in AP	1
Frequency, bandwidth, sub-carrier	3.5 GHz, 100MHz, 30kHz
Channel estimation, pilot size	MMSE, 24
UE transmission power	20 dBm
Path loss and channel fading	Ray tracing, Rayleigh fading
Noise figure, velocity of UEs	7 dB, {4, 30} km/h
User traffic	Full buffer, Uplink
Throughput requirements, $\tilde{\boldsymbol{g}}_{\boldsymbol{k}}$	{150, 200, 250} Mbps
Update intervals of AP cluster	100 msec
Time step length	50 msec
Initial AP cluster size $\boldsymbol{x}_{\boldsymbol{ini}}$	5
DRL parameters	
Online learning framework	Ape-X [14]
NN architecture for Ape-X	3 hidden layers with 512 units, 2 hidden layers [14]
Number of actors	100
Target network update intervals	2500
Network parameters copy intervals	500
Training batch size	512
Discount factor, learning rate	0.5, 0.00025/4
Episode length	200 time steps (20 seconds)
Design parameter $\boldsymbol{B}$, $\boldsymbol{O}$, $\boldsymbol{s}$	10, 400, 2
GA parameters	
Population size, number of generations, mutation rate	10, 50, 0.2

- Genetic Algorithm (GA): The GA is employed to identify the global minimum of non-linear optimization problems [15]. An individual is defined as the combination of the number of AP cluster sizes for each UE. The objective function is formulated as the cumulative reward for all UEs.
- Static Approach (SA): The size of the AP cluster for each UE is predetermined. The combination of APs in each AP cluster is selected based on signal strength, up to the pre-established AP cluster size. To meet the throughput requirements in SA, the AP cluster sizes are heuristically set to 10, 20, and 40 for UEs with throughput requirements of 150, 200, and 250 Mbps, respectively.
- Existing distributed DRL with a single learner (D-DRL): The D-DRL shows the results with a single learner based on all UE information, with the same DRL parameters as the proposed, and the DRL design is based on [11].

First, we show the simulation results of comparing the average throughput satisfaction rate versus the amount of training data in Fig. 3. The throughput satisfaction rate at a time step can be calculated as $\sum_{k\in K} q_k / K$. The proposed method achieves a higher throughput satisfaction rate with fewer episodes than existing D-DRL methods. This is because it learns efficiently even with less data and small NN models by using UEs in a similar radio environment, and the trained models infer for UEs with similar measurements and radio environments. On the other hand, the existing D-DRL with a single learner does not sufficiently learn the site-specific radio environment, resulting in low inference accuracy.

Next, Fig. 4 shows the average throughput satisfaction rate for the computational load of signal processing C^{comp}. The plot in the upper left indicates that C^{comp} is low and the throughput satisfaction rate is high. Here, the number of episodes for the proposed method and D-DRL is 3. The proposed method learns the site-specific radio environment with a small amount of learning data and achieves the best balance between throughput and computational load for signal processing. D-DRL has a low throughput satisfaction rate and C^{comp} because it under-allocates the AP cluster. SA has a high computational load because it over-allocates APs regardless of the radio environment. In GA, the search range is wide in this large-scale environment, so it does not obtain a high-quality sub-optimal solution, and neither throughput satisfaction nor C^{comp} matches the proposed method.

Fig. 5 shows the average computation time to determine the AP cluster for each UE and the number of parameters in NN. We use MATLAB and a personal computer (PC) with a Core i9-7900X, 64 GB of memory, and SSD storage. From Fig. 5, we find that the proposed method has the shortest computation time, less than the AP cluster period of 100 msec. This is because the proposed DRL design includes a smaller action space and NN model than D-DRL. The inference process of the proposed method can be parallelized for each actor, so it is possible to perform AP clustering that is sufficiently agile for user mobility even in larger environments. On the other hand, the computation time for the meta-heuristic optimization-based method using GA is long. This is because calculating the inverse matrix for each UE is necessary for calculating the objective function when repeating the search, which increases the computational load for the decision of AP clustering.

V. Conclusion

This paper proposed a distributed DRL method with multiple learners for AP clustering in CF-mMIMO. Using multiple learners, each associated with a distinct radio environment, the proposed method allows for efficient learning and improved inference accuracy, even with less data and small NN models. In addition, the proposed method dynamically adjusts the association between UE and learners according to the UE's movement, ensuring agility when changes in the radio environment occur. Simulation with an actual urban structure showed that the proposed method achieves better AP clustering with higher inference accuracy than existing methods, even with less learning data and smaller NN models.

Acknowledgments

These research results were obtained from the commissioned research (JPJ012368C00401) by National Institute of Information and Communications Technology (NICT), Japan.

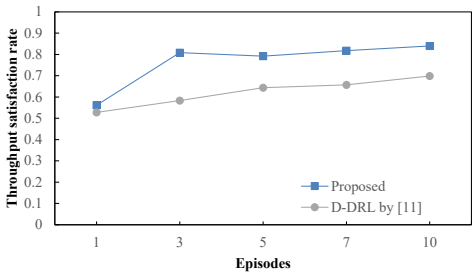

Fig. 3. Throughput satisfaction rate versus the amount of learning data.

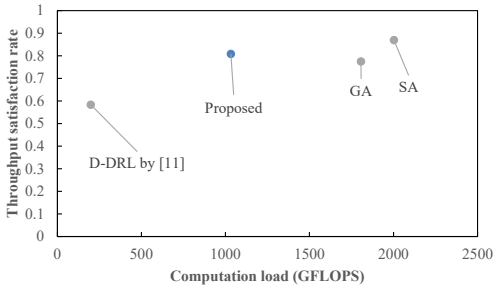

Fig. 4. Throughput satisfaction rate versus computational load for signal processing with different approaches.

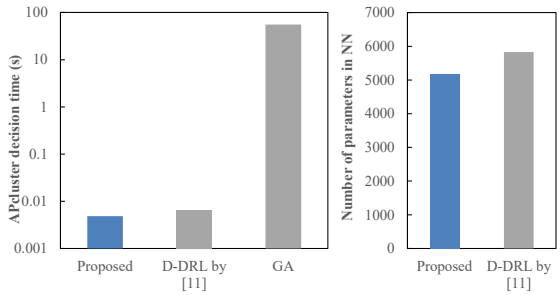

Fig. 5. Comparison for decision time for AP clustering and the number of parameters in NN.

References

[1] Next G Alliance, "6G Applications and Use Cases," Jun. 2022. https://www.nextgalliance.org/6g-library/

[2] The NGMN Alliance, "NGMN 6G USE CASES AND ANALYSIS," Feb. 2022. https://www.ngmn.org/

[3] H. Q. Ngo et al., "Cell-free massive MIMO: uniformly great service for everyone," in 2015 IEEE 16th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC), Jun. 2015, pp. 201–205.

[4] E. Bjornson et al., "Scalable cell-free massive MIMO systems," IEEE Trans. Commun., vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[5] C. D'Andrea et al., "User association in scalable cell-free massive MIMO systems," 54th Asilomar Conf. on Signals, 2020, pp. 826-830.

[6] A. Ikami et al., "Interference suppression for distributed CPU deployments in Cell-Free massive MIMO", the 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall).

[7] A. Ikami et al., "Cooperation Method Between CPUs in Large-Scale Cell-Free Massive MIMO for User-Centric RAN," in IEEE Access, vol. 11, pp. 95267-95277, 2023.

[8] X. Chai, et al. "Reinforcement Learning Based Antenna Selection in User-Centric Massive MIMO," in 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020.

[9] Y. Al-Eryani, et al. "Multiple Access in Cell-Free Networks: Outage Performance, Dynamic Clustering, and Deep Reinforcement Learning-Based Design," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 4, pp. 1028-1042, April 2021.

[10] N. Ghiasi, et al. "Energy Efficient AP Selection for Cell-Free Massive MIMO Systems: Deep Reinforcement Learning Approach," in IEEE Transactions on Green Communications and Networking, vol. 7, no. 1, pp. 29-41, March 2023.

[11] Y. Tsukamoto, et al. "User-centric AP Clustering with Deep Reinforcement Learning for Cell-Free Massive MIMO," In Proceedings of the Int'l ACM Symposium on Mobility Management and Wireless Access (MobiWac 2023).

[12] Z. Xiong, et al. "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," in IEEE Vehicular Technology Magazine, vol. 14, no. 2, pp. 44–52, Jun. 2019.

[13] O-RAN Alliance, "O-RAN: Towards an Open and Smart RAN," White Paper, Oct. 2018.

[14] D. Horgan, et al. "Distributed prioritized experience replay," Apr. 2018, arXiv:1803.00933.

[15] K. Goudos, "Evolutionary algorithms for wireless communications— a review of the state-of-the art," in Contemporary Issues in Wireless Commun. Rijeka, Croatia: InTech, 2014.