Multi-Agent Deep Reinforcement Learning for Dynamic Multi-AP Cooperation in Cell-Free Massive MIMO Networks

Mahnoor Ajmal, Sunghyun Kim, Ayesha Siddiqa, Youngjoon Yang, Deepak Singh, Dongkyun Kim
School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea
Emails: {mahnoor.ajmal, sunghyunkim, asiddiqa, youngj719, deepak.singh, dongkyun}@knu.ac.kr

***Abstract*—Cell-free massive MIMO (CF-mMIMO) emerges as a promising architecture for next-generation wireless networks, featuring geographically distributed access points (APs) that collectively serve users without traditional cell boundaries. However, it faces significant inter-AP interference in overlapping coverage areas, which degrades performance and limits spectral efficiency. Conventional solutions either employ computationally prohibitive centralized cooperation schemes or operate APs independently without cooperation, resulting in suboptimal interference mitigation. We introduce a novel framework utilizing multi-agent deep reinforcement learning (MADRL) to enable intelligent AP cooperation without requiring centralized control for interference mitigation. The framework incorporates a threshold-based cooperation mechanism that allows APs to autonomously identify high-impact cooperation opportunities while avoiding unnecessary resource consumption. Each AP deploys an independent Double Deep Q-Network (DDQN) agent that dynamically learns optimal cooperation strategies through environmental interaction and interference-aware reward mechanisms. Extensive simulation validates the efficacy of this method, which achieves an average SINR of 9.48 dB and a sum rate of 134.6 bps/Hz while utilizing 61.2% fewer cooperation links compared to centralized benchmarks. The intelligent cooperation strategy yields 2.73× higher efficiency than exhaustive cooperation methods while maintaining acceptable user fairness across the network. These results establish MADRL as a practical and effective solution for enabling intelligent interference mitigation in future generation wireless networks.**

***Index Terms*—massive MIMO, Cell Free, Interference Management, Multi Agent DRL, DDQN**

I. Introduction

The relentless demand for ubiquitous high-speed connectivity has driven the evolution of wireless network architectures beyond traditional cellular boundaries. Cell-Free massive Multiple-Input Multiple-Output (CF-mMIMO) has emerged as a promising technology that achieves this by deploying a distributed network of access points (APs) that cooperatively serve users. This architectural shift provides seamless connectivity and collectively enhances both spectral efficiency and coverage uniformity across geographical regions [1], [2].

However, a critical challenge in CF-mMIMO systems is managing the inherent inter-AP interference, which becomes particularly severe in dense network deployments. A user receives a desired signal from its serving APs while simultaneously experiencing interference from all other transmitting APs. This is most problematic in regions where AP coverage areas overlap, creating interference-limited zones that compromise system performance and user experience.

To address this interference challenge, CF-mMIMO systems require cooperation mechanisms that balance performance optimization with practical implementation constraints. Current implementation strategies have evolved into two primary approaches: centralized architectures where a Central Processing Unit (CPU) handles all cooperation decisions, and distributed architectures where individual APs make local optimization decisions independently. Each approach presents distinct tradeoffs between interference management effectiveness and system scalability.

Centralized architectures employ a CPU to handle all signal processing tasks, including user cluster formation, resource allocation, and precoding computation, while APs function primarily as relay nodes implementing CPU-generated decisions [3], [4]. While this approach achieves superior interference management through global network visibility, it suffers from computational bottlenecks that scale poorly with network size, fronthaul bandwidth constraints, and single points of failure that compromise system resilience.

Conversely, distributed architectures shift processing responsibilities to individual APs, with the CPU handling only data routing and high-level cooperation [5], [6]. This approach addresses scalability concerns but introduces a critical cooperation gap: each AP optimizes locally while remaining oblivious to its interference impact on neighboring APs. Traditional precoding techniques, such as minimum mean square error (MMSE), effectively mitigate intra-AP interference among coserved users but leave inter-AP interference fundamentally unaddressed [7]–[9].

The inter-AP interference challenge intensifies in dense CF-mMIMO deployments where AP coverage areas frequently overlap, as illustrated in Fig. 1. Without direct AP-AP cooperation mechanisms, interfering transmissions from non-serving APs can severely degrade user experience. This unmanaged interference represents a critical performance bottleneck that existing distributed architectures fail to address effectively.

Recent developments in machine learning, especially deep reinforcement learning (DRL), demonstrate potential for adaptive network optimization. Nevertheless, current ML-based methods generally address individual challenges like power control, energy efficiency, or user association [10]–[12], cre-

ating a research gap regarding intelligent AP cooperation.

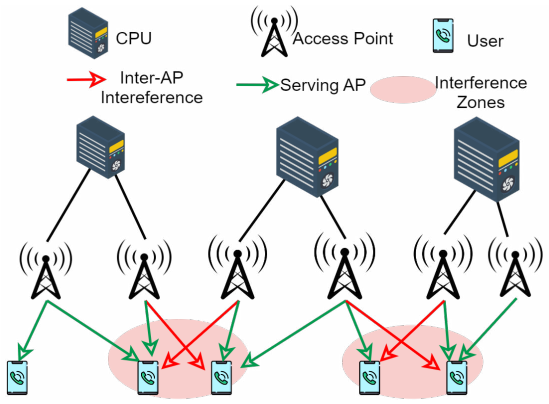

Fig. 1. Cell Free Massive MIMO Setup

Current research on AP cooperation in CF-mMIMO systems typically adopts an all-or-nothing approach: either full centralized control or complete AP independence. Although some studies mention AP-to-AP cooperation, they do not provide a practical framework for its implementation, instead assuming it happens automatically. These non-selective strategies generate substantial overhead because interference levels are highly variable and dynamic. Our core insight is that mitigating all interference isn't necessary. In many scenarios, interference is naturally low, or the effort required to reduce it is greater than the resulting performance benefit. Consequently, we introduce a threshold-based selective AP cooperation method, enabling APs to cooperate only when interference surpasses a specific limit. This strategy significantly reduces system overhead while focusing interference control precisely where it's most impactful.

We present a novel framework for selective AP collaboration in distributed CF-mMIMO systems using multi-agent deep reinforcement learning (MADRL). The proposed framework enables direct, autonomous AP-to-AP cooperation, bypassing CPU mediation and directly addressing the scalability-performance trade-off that limits current architectures. Each AP functions as an independent learning agent, continuously monitoring its local interference environment to make intelligent, adaptive cooperation decisions based on learned policies. This approach, centered on a threshold-adaptive cooperation protocol, offers several key contributions:

- **Autonomous & Adaptive cooperation**: Each AP independently learns and dynamically adjusts its cooperation strategy based on local observations, eliminating centralized bottlenecks.
- **Interference-Aware Selectivity**: Cooperation is initiated solely when interference mitigation benefits outweigh communication overheads, optimizing network resource utilization.
- **Scalable Distributed Architecture**: Enables practical deployment in large-scale networks via direct AP-AP communication, preserving cooperation effectiveness without reliance on centralized control.

The organization of this manuscript is outlined below: Section II details the system model and problem setup. Section III describes the proposed framework, while Section IV evaluates performance. The study concludes in Section V.

II. System Model and Problem Formulation

Consider a distributed CF-mMIMO system with M APs, each equipped with L antennas, serving K single-antenna users. The system operates using time-division duplex with coherence block length $\tau_c = \tau_p + \tau_u + \tau_d$, allocating τ_p and τ_u symbols for uplink pilot and data transmission and τ_d symbols for downlink data communication.

A. Channel Modeling and User-Centric Clustering

Using the composite fading model, we define the complex channel vector connecting AP m to user k as follows:

$$\mathbf{h}_{m,k} = \sqrt{\beta_{m,k}}\mathbf{g}_{m,k} \tag{1}$$

where $\beta_{m,k}$ captures large-scale propagation effects including path loss and shadowing, while $\mathbf{g}_{m,k} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$ represents small-scale fading between AP m and user k.

To maintain scalability, we employ user-centric clustering, in which a subset of APs $\mathcal{M}_k$ with the strongest channel serve each user k [5]. The serving AP set for user k is defined as:

$$\mathcal{M}_k = \{m : \beta_{m,k} \geq \delta \cdot \max_{m'=1,\ldots,M} \beta_{m',k}\} \tag{2}$$

where $\delta \in (0, 1]$ is a threshold parameter controlling cluster size. Correspondingly, the user set served by AP m serves user subset $\mathcal{K}_m = \{k : m \in \mathcal{M}_k\}$.

B. Channel Estimation

In the uplink training phase, each user k transmits pilot symbols $\boldsymbol{\phi}_k \in \mathbb{C}^{\tau_p}$ with power ρ_p. The received pilot signal at AP m m from all users is:

$$\mathbf{Y}_m = \sqrt{\rho_p}\sum_{k=1}^{K}\mathbf{h}_{m,k}\boldsymbol{\phi}_k^T + \mathbf{N}_m \tag{3}$$

where $\mathbf{N}_m \in \mathbb{C}^{L\times\tau_p}$ contains additive white gaussian noise samples. The MMSE channel estimator at AP m yields [13]:

$$\hat{\mathbf{h}}_{m,k} = \frac{\sqrt{\rho_p\tau_p}\beta_{m,k}}{\rho_p\tau_p\sum_{j\in\mathcal{K}}\beta_{m,j}|\boldsymbol{\phi}_j^H\boldsymbol{\phi}_k|^2 + \sigma^2}\mathbf{y}_{m,k} \tag{4}$$

Based on $\hat{\mathbf{h}}_{m,k}$, AP design precoding vector $\mathbf{w}_{m,k}$ and allocate power for downlink communication.

C. Downlink Signal Model and Interference Characterization

During downlink transmission, each AP m forms a composite signal for its served users:

$$\mathbf{x}_m = \sum_{k\in\mathcal{K}_m}\sqrt{\rho_{m,k}}\mathbf{w}_{m,k}s_k \tag{5}$$

where $\rho_{m,k}$ is transmit power, $\mathbf{w}_{m,k}$ is the normalized precoding vector, and s_k represents the data symbol for user k with

$\mathbb{E}[|s_k|^2] = 1$. The received signal at user k is the superposition of signals from multiple APs:

$$y_k = \underbrace{\sum_{m \in \mathcal{M}_k} \sqrt{\rho_{m,k}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,k} s_k}_{\text{desired signal}} + \underbrace{\sum_{m \in \mathcal{M}_k} \sum_{\substack{j \in \mathcal{K}_m \\ j \neq k}} \sqrt{\rho_{m,j}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,j} s_j}_{\text{intra-AP interference}} + \underbrace{\sum_{m \notin \mathcal{M}_k} \sum_{j \in \mathcal{K}_m} \sqrt{\rho_{m,j}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,j} s_j}_{\text{inter-AP interference}} + n_k \quad (6)$$

where $n_k \sim \mathcal{CN}(0, \sigma^2)$ represents additive noise.

A critical insight is that inter-AP interference originates from neighboring APs not serving a particular user, k. To quantify interference coupling between APs, we define the average interference power that AP n causes to users served by AP m as:

$$I_{m,n} = \frac{1}{|\mathcal{K}_m|} \sum_{k \in \mathcal{K}_m} \left(\sum_{j \in \mathcal{K}_n} \rho_{n,j} |\mathbf{h}_{n,k}^H \mathbf{w}_{n,j}|^2 \right) \quad (7)$$

This interference metric enables APs to assess the mutual impact of their transmission strategies and forms the foundation for intelligent cooperation decisions in our proposed framework.

D. Problem Formulation

Driven by the need to manage interference in CF networks, our objective is to develop an intelligent AP cooperation framework that optimizes system performance and minimizes coordination overhead. We formulate this as a multi-objective optimization problem:

$$\max_{\{\pi_m\}} \mathbb{E}\left[\sum_{k=1}^{K} \log_2(1 + \text{SINR}_k) \right] - \lambda \sum_{m=1}^{M} \sum_{n \in \mathcal{N}_m} a_{m,n} \quad (8)$$

where π_m represents AP m's cooperation policy, $\mathcal{N}_m$ denotes neighboring APs with significant interference coupling, and $a_{m,n}$ is the binary cooperation indicator incurring cooperation cost λ. The signal-to-interference-plus-noise ratio (SINR) for user k is:

$$\text{SINR}_k = \frac{|\sum_{m \in \mathcal{M}_k} \sqrt{\rho_{m,k}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,k}|^2}{\sum_{j \neq k} |\sum_{m \in \mathcal{M}_j} \sqrt{\rho_{m,j}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,j}|^2 + \sigma^2} \quad (9)$$

III. Proposed Multi-Agent Deep Reinforcement Learning Framework

This section introduces the proposed MADRL methodology designed to address inter-AP interference challenges through multi-AP cooperation. The developed MADRL approach establishes a natural correspondence with the distributed CF-mMIMO network topology, wherein individual APs operate as independent intelligent agents. We first outline the agent design fundamentals, including state representation, action spaces, and reward mechanisms, followed by the MADRL-Double Deep Q-Network (DDQN) learning framework implementation.

1) Agent Design and State Representation: Each AP m acts as an independent agent. At each time step t, agent m observes a local state vector $\mathbf{s}_m$:

$$\mathbf{s}_m = [\boldsymbol{\beta}_m, \mathbf{I}_m, \mathbf{C}_m, \mathbf{T}_m] \quad (10)$$

where $\boldsymbol{\beta}_m = \{\beta_{m,k} : k \in \mathcal{K}_m\}$ represents the large-scale fading coefficients between AP m and served users. $\mathbf{I}_m = \{I_{m,n_1}, I_{m,n_2}, \ldots\}$ represents normalized interference level vector, $\mathbf{C}_m$ represents current binary vector indicating cooperation status with neighbors, and $\mathbf{T}_m$ represents historical cooperation effectiveness.

2) Action Space and Neighbor Selection: For computational tractability, each AP considers cooperation only with high-interference neighbors:

$$\mathcal{N}_m = \{n : I_{m,n} > \gamma_{\text{th}}\} \quad (11)$$

Agent m's action $\mathbf{a}_m \in \{0,1\}^{|\mathcal{N}_m|}$ is a binary vector, where $a_{m,n} = 1$ signifies cooperation with AP n. The total number of distinct actions is $2^{|\mathcal{N}_m|}$, representing all possible cooperation combinations within its neighbor set, from which the agent selects one action per time step. The cooperation decisions determine cluster formation where cooperating APs $\{m, n : a_{m,n} = 1\}$ jointly process their common users.

3) Reward Function: The reward function r_m for AP m balances performance improvement against cooperation overheads:

$$r_m = \alpha_1 \frac{\sum_{k \in \mathcal{K}_m} R_k^{\text{coop}} - R_k^{\text{baseline}}}{\sum_{k \in \mathcal{K}_m} R_k^{\text{baseline}}} - \alpha_2 \sum_{n \in \mathcal{N}_m} a_{m,n} - \alpha_3 \sum_{n \in \mathcal{N}_m} I_{m,n} \cdot a_{m,n} \quad (12)$$

where R_k^{coop} and R_k^{baseline} are rates with and without cooperation. The third term encourages the agent to prioritize forming partnerships that tackle the most severe interference, maximizing the marginal utility of cooperation, and coefficients $\alpha_1, \alpha_2, \alpha_3$ serve as tunable weights.

A. Double Deep Q-Network Implementation

Each AP implements an independent DDQN agent for learning cooperation policies. We selected DDQN to mitigate the overestimation bias present in standard DQN, while leveraging its dual-network structure for enhanced data efficiency and stable exploration. The online Q-network (Q_{main}) handles action selection, whereas the target Q-network (Q_{target}) provides stable value estimates. Our implementation employs three fully connected layers (128-64-32 neurons) with ReLU activations and dropout regularization. Experience replay buffers are utilized to improve sample efficiency.

The DDQN target computation for bias reduction follows:

$$Y_j = r_j + \gamma Q_{\text{target}}\left(\mathbf{s}_{j+1}, \arg\max_{\mathbf{a}'} Q_{\text{main}}(\mathbf{s}_{j+1}, \mathbf{a}'; \theta), \theta_{\text{target}}\right) \quad (13)$$

where Q_{main} selects actions while Q_{target} evaluates their values, effectively decoupling these operations to reduce overestimation. We optimize the online network weights using MSE loss with Adam optimizer against these computed targets. The target network parameters are updated periodically to maintain training stability.The DDQN architecture is shown in Fig. 2.

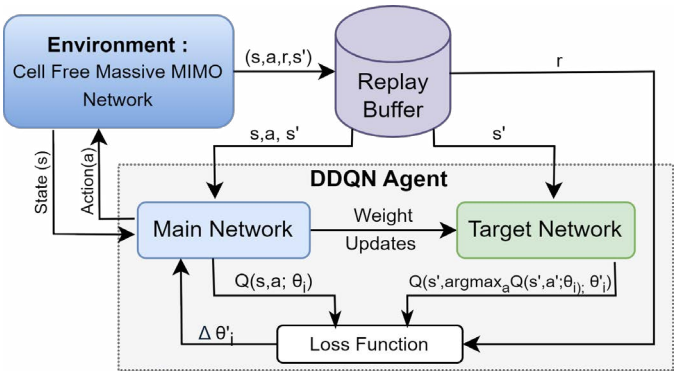

Fig. 2. DDQN Architecture

B. Cooperative Precoding

When APs decide to cooperate based on their learned policies, they form dynamic clusters for joint interference management. A cooperative cluster $\mathcal{C}$ consists of a set of APs that have mutually agreed to cooperate. APs not in any such cluster operate independently. For a cooperative cluster $\mathcal{C}$, the member APs perform joint regularized zero-forcing (RZF) precoding to suppress inter-cluster interference for the set of all users they serve, $\mathcal{K}_{\mathcal{C}} = \bigcup_{m \in \mathcal{C}} \mathcal{K}_m$.

First, for each user $k \in \mathcal{K}_{\mathcal{C}}$, the joint channel vector from the cluster is constructed by vertically stacking the individual AP-user channel vectors:

$$\mathbf{h}_k^{\mathcal{C}} = [\mathbf{h}_{m_1,k}^T, \mathbf{h}_{m_2,k}^T, \ldots, \mathbf{h}_{m_{|\mathcal{C}|},k}^T]^T \in \mathbb{C}^{|\mathcal{C}|L \times 1} \tag{14}$$

where $\{m_1, \ldots, m_{|\mathcal{C}|}\}$ is the set of APs in $\mathcal{C}$. The composite channel matrix for the entire cluster, $\mathbf{H}_{\mathcal{C}}$, is then formed by horizontally concatenating the joint channel vectors of all users in $\mathcal{K}_{\mathcal{C}}$:

$$\mathbf{H}_{\mathcal{C}} = [\mathbf{h}_{k_1}^{\mathcal{C}}, \mathbf{h}_{k_2}^{\mathcal{C}}, \ldots, \mathbf{h}_{k_{|\mathcal{K}_{\mathcal{C}}|}}^{\mathcal{C}}] \in \mathbb{C}^{|\mathcal{C}|L \times |\mathcal{K}_{\mathcal{C}}|} \tag{15}$$

The joint RZF precoding matrix for the cluster is then computed as:

$$\mathbf{W}_{\mathcal{C}} = \mathbf{H}_{\mathcal{C}}^H (\mathbf{H}_{\mathcal{C}} \mathbf{H}_{\mathcal{C}}^H + \alpha \mathbf{I})^{-1} \tag{16}$$

where $\alpha > 0$ is the regularization parameter. The j-th column of $\mathbf{W}_{\mathcal{C}}$, denoted $\mathbf{w}_{k_j}^{\mathcal{C}}$, represents the joint precoding vector for user k_j.

To find the individual precoding vector $\mathbf{w}_{m,k}$ for user k at a specific AP $m \in \mathcal{C}$, we extract the relevant L-dimensional segment from the joint vector $\mathbf{w}_k^{\mathcal{C}}$. To do this robustly, we define $\text{idx}(m, \mathcal{C})$ as the 1-indexed position of AP m within a predetermined, ordered list of the APs in $\mathcal{C}$. The block of rows corresponding to AP m can then be identified and extracted. The final normalized precoding vector is:

$$\mathbf{w}_{m,k} = \frac{[\mathbf{w}_k^{\mathcal{C}}]_{L(\text{idx}(m,\mathcal{C})-1)+1:L\cdot\text{idx}(m,\mathcal{C})}}{\|[\mathbf{w}_k^{\mathcal{C}}]_{L(\text{idx}(m,\mathcal{C})-1)+1:L\cdot\text{idx}(m,\mathcal{C})}\|} \tag{17}$$

For non-cooperating APs (i.e., those in a cluster of size one), standard maximum ratio transmission (MRT) precoding is employed as a baseline:

$$\mathbf{w}_{m,k} = \frac{\hat{\mathbf{h}}_{m,k}}{\|\hat{\mathbf{h}}_{m,k}\|} \tag{18}$$

IV. Performance Evaluation

A. Simulation Setup

We evaluate the proposed MADRL framework through extensive MATLAB simulations. Table I summarizes the key system parameters, which are selected to represent realistic CF-mMIMO scenarios.

TABLE I
System Simulation Parameters

Parameter	Value
Number of APs (M)	100
Number of users (K)	40
Antennas per AP (L)	4
Coverage area	1000 m × 1000 m
Maximum serving APs per user	5
Maximum neighbors per AP	4
Interference threshold (γ_{th})	0.05 × max(I_{AP})
Coherence block length (τ_c)	200 symbols
Pilot training duration (τ_p)	10 symbols
AP transmit power	1000 mW
Noise figure	7 dB

B. Multi-Agent DRL Implementation

Each agent employs DDQN with architecture: input layer (state dimension), hidden layers (128→64→32 neurons with ReLU), and output layer (Q-values for all actions). Training employs an ϵ-greedy exploration strategy with adaptive decay, while target networks are updated every 10 episodes for learning stability. A summary of the associated configuration parameters is provided in Table II.

TABLE II
Multi-Agent DRL Configuration

DRL Parameter	Value/Setting
Agent Type	Double Deep Q-Network (DDQN)
Network Architecture	128-64-32 neurons
Activation Function	ReLU
Regularization	Dropout (0.1)
Experience Buffer Size	1000 transitions
Mini-batch Size	32
Target Update Frequency	10 episodes
Learning Rate	0.001
Discount Factor (γ)	0.95
Initial Exploration (ϵ)	0.8
Final Exploration (ϵ_{min})	0.01
Exploration Decay	0.005
Training Episodes	200

C. Benchmark Methods

We compare our proposed MADRL-based cooperation with two benchmarks: No cooperation, which involves traditional distributed processing without any direct AP-AP cooperation, and Centralized CPU-based full cooperation, where the CPU manages complete cooperation with global CSI for all APs.

D. Results and Discussion

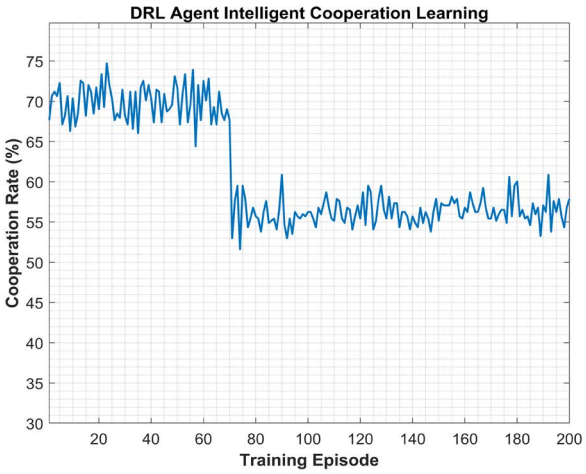

Fig. 3. DRL Training Convergence Analysis

Fig. 3 illustrates the evolution of cooperation rates throughout the 200-episode training process. The learning curve exhibits three distinct phases. An initial exploration phase (episodes 1–70) shows the rate fluctuating between 65–75% as agents explore various strategies. In the subsequent transition phase (episodes 71–100), the rate decreases to 50–60%. Finally, the convergence phase (episodes 101–200) demonstrates a stabilization of the cooperation rate at 55–58%. This pattern

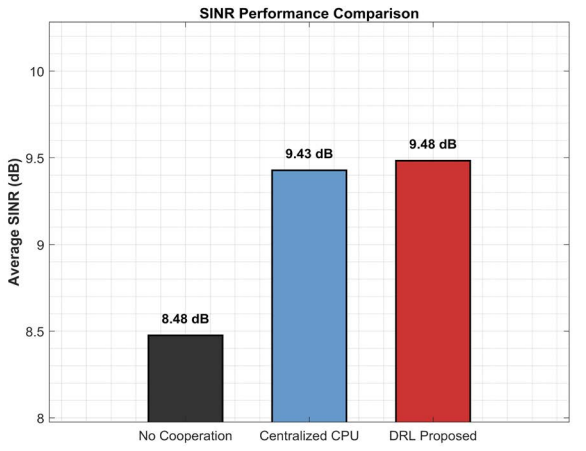

Fig. 4. SINR Performance Comparison

shows that agents learn to selectively cooperate, which yields better long-term rewards than indiscriminate cooperation.

Fig. 4 presents average SINR performance across three methods. No-cooperation achieves 8.48 dB baseline, centralized cooperation improves to 9.43 dB (0.95 dB gain), while DRL achieves 9.48 dB, slightly outperforming centralized methods by 0.05 dB in fully distributed operation. This demonstrates that intelligent cooperation can match centralized performance, with DRL's edge indicating learned policies may avoid interference-inducing decisions challenging for centralized controllers in dynamic environments.

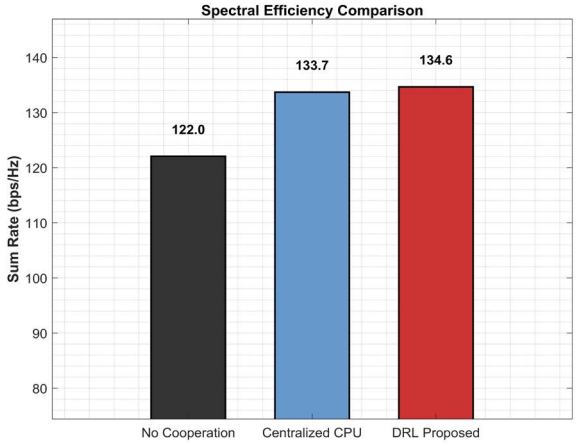

Fig. 5. Sum Rate Comparison

Fig. 5 examines sum rate performance for system spectral efficiency. No-cooperation yields 122.0 bps/Hz, centralized cooperation achieves 133.7 bps/Hz, while DRL attains 134.6 bps/Hz, representing 10.3% improvement over baseline. Centralized cooperation requires 500 links (efficiency: 0.023 bps/Hz/link) while DRL achieves superior performance with only 194 links (efficiency: 0.065 bps/Hz/link). This 2.8× efficiency advantage shows DRL learns high-impact cooperation opportunities while avoiding redundant links.

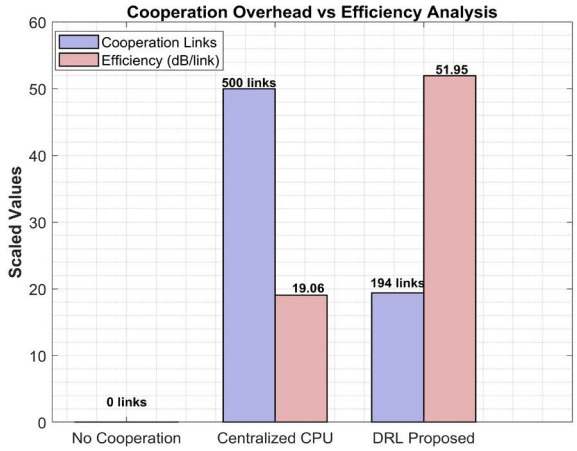

Fig. 6. Cooperation Overhead vs Efficiency Trade-off

Fig. 6 highlights the trade-off between cooperation overhead and system efficiency. The grouped bar chart illustrates the differences in resource utilization between centralized and DRL approaches. Centralized cooperation establishes 500 cooperation links with an efficiency of 19.06 scaled units, while DRL achieves 51.95 scaled units with only 194 links. This visualization demonstrates the core advantage of the proposed approach: DRL learns to achieve superior performance through intelligent selectivity rather than exhaustive cooperation. The 61.2% reduction in cooperation links translates directly to reduced fronthaul bandwidth requirements, lower computational complexity at central processing units, and improved system scalability. The DRL approach essentially discovers that strategic, targeted cooperation can outperform brute-force cooperation strategies.

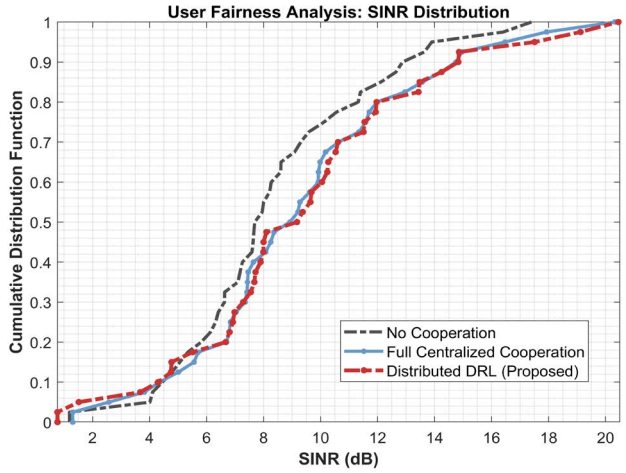

Fig. 7. User Fairness and Distribution Analysis

Fig. 7 presents user SINR cumulative distribution, revealing fairness characteristics across cooperation strategies. No cooperation (dashed black) shows dispersed distribution with significant low-SINR users, while centralized cooperation (solid blue) improves fairness by shifting toward higher values. DRL (dashed red) closely matches centralized performance, particularly in the critical 8-12 dB range where most users operate, maintaining satisfactory service quality without degrading worst-case performance. This comprehensive evaluation validates the proposed MADRL approach effectiveness.

V. Conclusion

We proposed a multi-agent deep reinforcement learning approach for AP cooperation in cell-free massive MIMO networks. Our method utilizes distributed DDQN agents to make selective cooperation decisions, effectively managing inter-AP interference without centralized control. Simulation results show significant performance gains, achieving a 9.48 dB average SINR and a 134.6 bps/Hz sum rate with 61.2% fewer cooperation links than centralized schemes. This translates to 2.73× better efficiency, demonstrating that intelligent cooperation selection outperforms both non-cooperative operation and comprehensive centralized coordination. The key insight is that not all cooperation is beneficial - our DDQN agents learn to identify high-impact cooperation opportunities while avoiding redundant links. Future work will investigate continuous cooperation strategies using advanced DRL algorithms with a continuous action space for even finer-grained interference control in dense networks.

VI. Acknowledgment

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT), (NRF-2022R1A2C1003620). Additionally, this work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-RS-2018-NR031059

References

[1] J. Zheng, J. Zhang, H. Du, D. Niyato, B. Ai, M. Debbah, and K. B. Letaief, "Mobile cell-free massive mimo: Challenges, solutions, and future directions," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 140–147, 2024.

[2] M. Ajmal, A. Siddiqa, B. Jeong, J. Seo, and D. Kim, "Cell-free massive multiple-input multiple-output challenges and opportunities: A survey," *ICT Express*, vol. 10, no. 1, pp. 194–212, 2024.

[3] A. Lancho, G. Durisi, and L. Sanguinetti, "Cell-free massive mimo for urllc: A finite-blocklength analysis," *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 8723–8735, 2023.

[4] D. Maryopi, M. Bashar, and A. Burr, "On the uplink throughput of zero forcing in cell-free massive mimo with coarse quantization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7220–7224, 2019.

[5] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.

[6] M. Ajmal, M. A. Tariq, M. M. Saad, S. Kim, and D. Kim, "Scalable cell-free massive mimo networks using resource-optimized backhaul and pso-driven fronthaul clustering," *IEEE Transactions on Vehicular Technology*, 2024.

[7] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSEO processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2020.

[8] G. Interdonato, M. Karlsson, E. Bjornson, and E. G. Larsson, "Downlink spectral efficiency of cell-free massive MIMO with full-pilot zero-forcing," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1003–1007, IEEE, 2018.

[9] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4758–4774, 2020.

[10] Z. Gao, Q. Zhang, J. Liu, Z. Du, and Y. Li, "Drl-based ap selection in downlink cell-free massive mimo network with pilot contamination," *IEEE Communications Letters*, 2024.

[11] Z. Wu, Y. Jiang, Y. Huang, F.-C. Zheng, and P. Zhu, "Energy-efficient joint ap selection and power control in cell-free massive mimo systems: A hybrid action space-drl approach," *IEEE Communications Letters*, 2024.

[12] L. Sun, J. Hou, and R. Chapman, "Multi-agent deep reinforcement learning for access point activation strategy in cell-free massive mimo networks," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, IEEE, 2023.

[13] E. Björnson and L. Sanguinetti, "Making cell-free massive mimo competitive with mmse processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2019.