Converging Game Theory and Reinforcement Learning For Industrial Internet-of-Things

Tai Manh Ho, Kim-Khoa Nguyen, and Mohamed Cheriet

***Abstract*—The fifth-generation (5G) wireless network provides high-rate, ultra-low latency, and high-reliability connections that can meet the Industrial Internet of Things (IIoT) requirements in factory automation, especially for robot motion control. In this paper, we address 5G service provisioning in an automated warehouse scenario, where swarm robotics is controlled by an industrial controller that provides routing and job instructions over the 5G network. Leveraging the coordinated multipoint (CoMP), we formulate a time-varying joint CoMP clustering and 5G ultra-reliable low-latency communication (URLLC) beamforming design problem to control the robots that move around the automated warehouse for goods storage with the planned reference tracks. Traditional iterative optimization approaches are impractical in such a dynamic wireless environment due to high computational time. We propose a game-theoretic CoMP clustering algorithm combined with the Proximal Policy Optimization method to obtain a stationary solution closed to that of the exhaustive search algorithm considered as the global optimal solution.**

***Index Terms*—5G network, industrial IoT, URLLC**

I. Introduction

The "Fourth Industrial Revolution" is considered the automation revolution thanks to the innovations of 5G wireless communications, automation technologies, and artificial intelligence. Ultra-reliable and low-latency communication (URLLC) service provided by the 5G wireless network is able to fulfill the stringent requirement of factory automation, e.g. 10^{-9} packet loss probability and 99.9999% availability in motion control and mobile robot use cases [1]. However, guaranteeing extremely high reliability is challenging in such a dynamic environment of an automated warehouse with high mobility automated guided vehicles (AGV). Coordinated Multi-Point (CoMP) communication technique [2] that leverages spatial diversity is promising to achieve URLLC by sending duplicate data streams over diverse paths [3]. In the automated warehouse scenario, CoMP can combine the signals from multiple radio base stations (gNBs) so that highly dependable communications can be achieved to the moving objects, i.e., AGVs with the physical obstructions, e.g. warehouse racks and shelves.

Along with the advantages that CoMP can bring to the wireless network, providing CoMP-enabled URLLC wireless communication in the industrial Internet of Things (IIoT) is especially challenging due to highly dynamic radio frequency variations from moving objects (such as AGVs) [3]. Therefore, designing a joint CoMP clustering and beamforming for transmission between gNBs and AGVs that satisfies the URLLC constraints becomes more difficult and significantly different from that in conventional communication systems.

A. Prior Works

During the past few years, plenty of works try to coordinate CoMP transmission to improve the URLLC service in the 5G network via spatial diversity. In [4], Nasir *et al.* develop path-following algorithms, which generate a sequence of improved feasible points to solve the problem of resource allocation and beamforming design in the short blocklength regime for URLLC. In [5], Yang *et al.* formulate the CoMP-enabled RAN slicing problem for multicast enhanced mobile broadband (eMBB) and bursty URLLC service multiplexing as a multi-timescale optimization problem with a goal of maximizing eMBB and URLLC slice utilities, subject to total system bandwidth and transmit power constraints. In [6], Khan *et al.* propose a novel packet delivery mechanism, queuing strategy, and time-frequency resource allocation for CoMP-enabled URLLC in C-RAN architecture. In [7], the authors investigate the CoMP-enabled RAN slicing for bursty URLLC and eMBB service provision by deriving the minimum upper bound of network bandwidth orchestrated for URLLC traffic transmission to guarantee the URLLC packet blocking probability. In [8], the authors propose a heuristic resource allocation algorithm for CoMP-enabled URLLC with short packet communication by maximizing the availability of the CoMP. In [9], the authors propose to use an alternating direction method of multipliers (ADMM) for solving the resource optimization problem of the CoMP-enabled RAN slicing for massive Internet of things (mIoT) and bursty URLLC service multiplexing.

To overcome the shortcomings of traditional optimization theory, recent works have proposed to use of deep reinforcement learning (DRL) to address important aspects of CoMP communication such as clustering and beamforming design. In [10], a hybrid DRL model combining a deep deterministic policy gradient (DDPG) and a deep double Q-network (DDQN) model is proposed to cluster the access points and optimize the beamforming vectors to maximize the sum rate. The authors in [11] propose a deep Q-network (DQN)-based algorithm for jointly optimizing beamforming,

This work was supported by Mitacs, Ciena, and ENCQOR under Grant IT13947. (Corresponding author: Tai Manh Ho)

The authors are with Synchromedia Lab, École de Technologie Supérieure, Université du Québec, QC, Canada (email: manh-tai.ho.1@ens.etsmtl.ca;kim-khoa.nguyen@etsmtl.ca;mohamed.cheriet@etsmtl.ca).

power control, and interference coordination for voice bearers and data bearers in sub-6 GHz and millimeter-wave in 5G wireless network. In [12] the authors propose a distributed dynamic downlink-beamforming coordination algorithm based on the DQN method to improve the system capacity of this multi-cell multi-input single-output (MISO) interference channel. In [13], a multi-agent RL-based method is proposed for solving the problem of user-centric transmission/reception point (TRP)-grouping and user-association in joint transmission aided coordinated multipoint (CoMP) technique.

The power of game theory in solving many engineering problems has been proven. Therefore, combining reinforcement learning and game theory has recently attracted the attention of scholars [14], [15]. Shi *et al.* [14] propose a combination of the mean-field game (MFG) and DRL in which a DRL agent learns with the guidance of the Nash equilibrium solved by the MFG. The trust region policy optimization (TRPO) is applied to obtain the optimal solution to the problem modeled by MFG in [15]. Different from the existing works considering the combination of DRL and game theory, we propose a distributed framework in which the players of the game (i.e., AGVs) use the actions of the agents of the DRL (i.e., gNBs) to obtain a Nash equilibrium. In turn, the output of the game, i.e., the Nash equilibrium, is used as a network state to train the agents of the DRL model.

B. Motivation and Contribution

Most existing works which employ traditional iterative optimization approaches are unable to handle the time-varying dynamic environment with high mobility of the network entities which is the case in this paper. Traditional approaches can guarantee convergence to a locally optimal solution at the cost of complexity and computation time, which is not compatible with mission-critical applications. To the best of our knowledge, this is the first work that combines DRL and game theory to solve the high complexity problem of joint beamforming design and CoMP clustering in IIoT. In this paper, we propose a distributed low complexity game-theoretic CoMP clustering algorithm combined with the Proximal Policy Optimization (PPO) method to obtain an optimal solution for beamforming design for URLLC transmission between the gNBs and the AGVs in a highly dynamic environment of an automated warehouse application presented in Fig. 1. The main contributions of this paper can be summarized as follows:

- We formulate the time-varying problem of joint CoMP clustering and beamforming design for 5G URLLC transmission in industrial automation applications. The wireless channel condition is highly dynamic due to the high mobility of the AGVs in an automated warehouse scenario. Therefore, the traditional optimization approach is unable to handle the formulated problem in such a dynamic environment.
- We propose a multi-agent Proximal Policy Optimization (PPO) based algorithm to obtain an optimal policy of the beamforming design for the transmission of the gNBs.
- We then propose a low complexity game-theoretic CoMP clustering algorithm that uses the actions of the multi-agent PPO-based algorithm to obtain a Nash equilibrium of the formulated CoMP clustering game among AGVs. In turn, the Nash equilibrium of the CoMP clustering game will be used as a system state to train the agents of the PPO-based algorithm.
- The intensive simulation results demonstrate the effectiveness of our proposed framework in handling the interference caused by the increasing number of AGVs in the network.

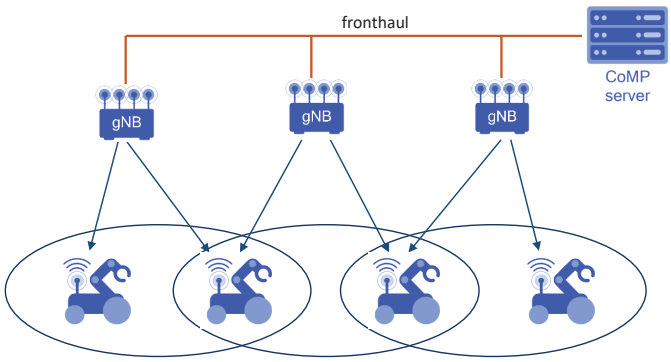


Figure 1: CoMP in factory automation.

The rest of the paper is organized as follows: Section II presents the system model and problem formulation. Section III presents an approach for user-centric CoMP clustering and the problem transformation. Section IV introduces the Proximal Policy Optimization algorithm followed by Section V presents a game theoretic approach for CoMP clustering. Section VI presents simulation results. Finally, Section VII concludes the paper.

II. System Model and Problem Formulation

We consider an automated warehouse IIoT network with a set of B radio base stations (gNodeBs or gNBs) denoted as $\mathcal{B}$, each gNB with M-antennas, and a set of K single-antenna AGVs denoted as $\mathcal{K}$. The AGVs move around the warehouse for goods storage with planned reference tracks (Fig. 2). Each AGV traces its planned reference track. Each AGV can be served by a set of $B_k[t] < B$ gNBs at time t. The set $\mathcal{B}_k \subset \mathcal{B}$ consisting of B_k gNBs is the CoMP cluster of AGV k, represents the minimum number of gNBs which can provide 5G communications with the required reliability to AGV k. Note that, these CoMP clusters can be overlapped in which a gNB can be in different clusters that serve different AGVs.

Moreover, we denote $\mathcal{K}_j \subset \mathcal{K}$ as the set of AGVs that are served by gNB j. All gNBs are connected to a single CoMP server over optical fiber fronthaul links. The CoMP enables the distributed gNBs to collaborate and simultaneously serve all AGVs within the warehouse area. We assume all the gNBs are deployed on the ceiling of the warehouse.

This article has been accepted for publication in IEEE Transactions on Network and Service Management. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2022.3202168

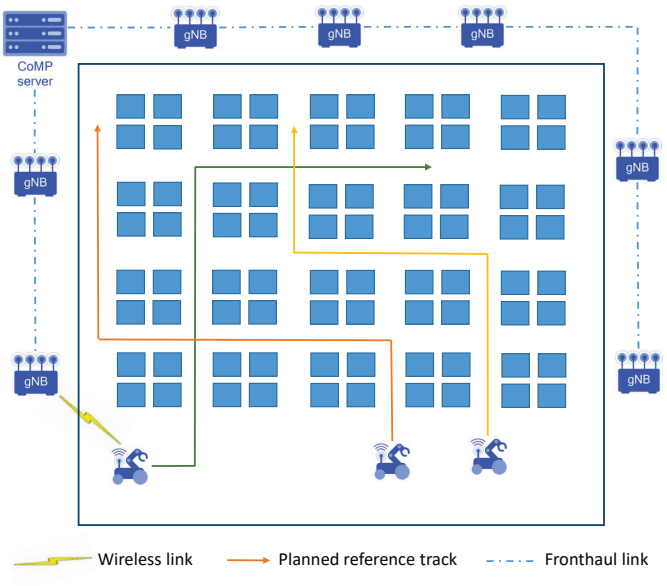


Figure 2: System model.

Let $\mathbf{q}_{\text{gNB},j} = [x_{\text{gNB},j}, y_{\text{gNB},j}]$ denotes the coordinate of the gNB j and $z_{\text{gNB},j}$ is the height of the gNB j.

The reference track is defined for each AGV k at each time step t as $X_k[t] = (\mathbf{q}_k[t], \theta_k[t])$ where $\mathbf{q}_k[t] = [x_k[t], y_k[t]]$ represents the spatial coordinates, and $\theta_k[t]$ is the orientation of the AGV. The control input $u_k[t] = \{v_k[t], \omega_k[t]\}$ sent from the controller implemented in the CoMP server to the k-th AGV consists of an intended translational velocity $v_k[t]$ and rotational velocity $\omega_k[t]$ at each time instant t. The AGV kinematic can be expressed as follows:

$$X_k[t+1] = X_k[t] + \Delta T \Theta_k[t] u_k[t], \tag{1}$$

where ΔT is the time slot duration and $\Theta_k[t]$ is given by:

$$\Theta_k[t] = \begin{bmatrix} \cos\theta_k[t] & 0 \\ \sin\theta_k[t] & 0 \\ 0 & 1 \end{bmatrix}. \tag{2}$$

The distance between the gNB j and AGV k at time instant t is

$$d_{k,j}[t] = \sqrt{\left\|\mathbf{q}_{\text{gNB},j} - \mathbf{q}_k[t]\right\|^2 + z^2_{\text{gNB},j}} \tag{3}$$

The real-time position of the AGVs can be tracked by 5G positioning techniques such as Downlink-Time Difference Of Arrival (DL-TDOA), Downlink-Angle Of Departure (DL-AoD), Uplink-Relative Time Of Arrival (UL-RTOA), Uplink-Angle of Arrival (UL-AoA), etc [16].

A. Communication Model

In reality, the channel state information (CSI) can be estimated by the CoMP through training the pilot sequences. Since the moving distance of an AGV in each time slot is substantially much smaller than the communication coverage of a gNB, we assume that CSI remains constant (fixed) within a slot but can vary across different time slots. Denote $\mathbf{w}_{k,j}$ as the transmit beamformer for the AGV k from the gNB j. Let s_k denote the complex data symbol for the AGV k and $\mathbb{E}\left[|\mathrm{s}_k|^2\right] = 1$, and $\sigma_k \sim \mathcal{CN}(0, \sigma_0^2)$ is the additive white Gaussian noise (AWGN) at the AGV k. The received signal y_k at AGV k can be expressed as[1]

$$\mathrm{y}_k = \underbrace{\sum_{j=1}^{B_k} \mathbf{h}_{k,j}^H \mathbf{w}_{k,j} \mathrm{s}_k}_{\text{Desired signal}} + \underbrace{\sum_{k' \neq k}^{K} \sum_{j=1}^{B_{k'}} \mathbf{h}_{k,j}^H \mathbf{w}_{k',j} \mathrm{s}_{k'}}_{\text{Interference}} + \sigma_k, \tag{4}$$

where $\mathbf{h}_{k,j} \in \mathbb{C}^{M \times 1}$ denotes the time-varying channel from the gNB j to the AGV k, and $\mathbf{h}_{k,j} = \sqrt{g_{k,j}} \tilde{\mathbf{h}}_{k,j}$ where $g_{k,j}$ accounts for the distance-based large-scale fading including path-loss component and shadow fading, and $\tilde{\mathbf{h}}_{k,j}$ is the small-scale fading vector associated with the channels between the gNB j and the AGV k. The large-scale fading channel gain $g_{k,j}$ between the gNB j and the AGV k can be expressed as:

$$g_{k,j} = \left(\frac{c}{4\pi f_c}\right)^2 \left(\frac{d_{k,j}}{d_0}\right)^{-\alpha_g}, \tag{5}$$

where f_c is the carrier frequency, c is the speed of light, $d_{k,j}$ is the distance between the gNB j and the AGV k, d_0 is a far field reference distance, and α_g is the path-loss exponent ($\alpha_g \in [2, 6]$). We assume the small-scale fading from the gNB and the AGV follows the Nakagami-m fading model [17]. The probability density function of random variable $\tilde{h}^{(l)}_{k,j} \in \tilde{\mathbf{h}}_{k,j}$, the small-scale fading channel gain between the l-th antenna of eNB j and AGV k, can be expressed as [17]:

$$f(z, m) = \frac{2m^m}{\Gamma(m)\Omega^m} z^{2m-1} \exp\left(-\frac{m}{\Omega} z^2\right), \tag{6}$$

where m is the fading parameter, $\Omega = \mathbb{E}\left[|\tilde{h}^{(l)}_{k,j}|^2\right]$, and $\Gamma(.)$ is the Gamma function. We assume that the CoMP server has knowledge of the instantaneous channel vectors $\{\mathbf{h}_{k,j}, \forall k \in \mathcal{K}, \forall j \in \mathcal{B}\}$.

The signal-to-interference-plus-noise ratio (SINR) and the Shannon achievable rate at the AGV k when using CoMP are given by [2]:

$$\gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \frac{\left|\sum_{j=1}^{B_k} \mathbf{h}_{k,j}^H \mathbf{w}_{k,j}\right|^2}{\sum_{k' \neq k}^{K} \left|\sum_{j=1}^{B_{k'}} \mathbf{h}_{k,j}^H \mathbf{w}_{k',j}\right|^2 + \sigma_k^2}, \tag{7}$$

$$\tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \log_2\Big(1 + \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j})\Big). \tag{8}$$

The maximum transmission rate to transmit D_k bits over n_k complex symbols in finite blocklength regime can be accurately approximated as [18]:

$$R_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) - \sqrt{\frac{V}{n_k}} \frac{Q^{-1}(\epsilon_k)}{\ln(2)} \geq \frac{D_k}{n_k}, \tag{9}$$

where ϵ_k is the decoding error probability, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$ and Q^{-1} is the inverse of Q.

[1] $\mathbf{x}^H$ is denoted the conjugate transpose operator.

The achievable decoding error probability of the AGV k in terms of γ_k and n_k can be expressed as follows:

$$\epsilon_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \leq Q\left(\frac{\ln(2)\left(\tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) - \frac{D_k}{n_k}\right)}{\sqrt{\frac{V}{n_k}}}\right). \tag{10}$$

where $V = 1 - \frac{1}{(1+\gamma_k)^2}$ is the channel dispersion. We assume that the packet size D_k and complex symbol n_k are the same for all gNBs in the set $\mathcal{B}_k$ corresponding to the AGV k.

From (10), it can be seen that, when the SINR $\gamma_k > 5$, the channel dispersion V can be accurately approximated by one $V \approx 1$, then the achievable decoding error probability can be rewritten as:

$$\epsilon_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \leq Q\left(\ln(2)\sqrt{n_k}\left(\tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) - \frac{D_k}{n_k}\right)\right). \tag{11}$$

In low SINR regime (i.e., $\gamma_k < 5$), equation (11) can be considered the upper bound of the decoding error probability.

According to (10) and (7), the mathematical expression of the required beamformers $\{\mathbf{w}_{k,j}\}$ from the gNBs to AGV k that satisfies the decoding error probability ϵ_k requirements can be written as

$$\gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k), \tag{12}$$

where the SINR threshold γ_k^{th} is defined as follows:

$$\gamma_k^{th}(\epsilon_k) = \exp\left(\frac{D_k \ln(2)}{n_k} + \sqrt{\frac{V}{n_k}} Q^{-1}(\epsilon_k)\right) - 1. \tag{13}$$

B. Problem Formulation

At each time instant, depending on the real-time position $\{X_k[t]\}$ of the AGVs, the CoMP server dynamically performs the CoMP clustering by assigning the set $\mathcal{B}_k$ gNBs from the available B gNBs to each AGV. Then the corresponding optimal beamforming vectors are computed so that the SINR γ_k of the AGV k meets the ultra-reliability requirement. We consider the joint problem of CoMP clustering and beamforming design with the objective of sum-rate maximization for all AGVs subject to the URLLC constraint. Specifically, the joint problem in time slot t can be formulated as follows:

P1A:

$$\max_{\{\mathcal{B}_k\},\{\mathbf{w}_{k,j}\}} \sum_{k\in\mathcal{K}} R_k(\mathbf{w}_{k,j}[t], \mathbf{h}_{k,j}[t]) \tag{14a}$$

$$\text{subject to: } \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k),\ \forall k \in \mathcal{K}, \tag{14b}$$

$$\sum_{k\in\mathcal{K}_j} \|\mathbf{w}_{k,j}[t]\|^2 \leq P_j,\ \forall j \in \mathcal{B}, \tag{14c}$$

$$\mathcal{B}_k[t] \subset \mathcal{B},\ \forall k \in \mathcal{K}. \tag{14d}$$

Constraint (14b) guarantees the reliability communication of AGV k, whereas (14c) sets a constraint on the total transmit power of gNB j. It can be seen that the problem **P1A** in (14) is non-convex combinatorial due to the non-convex objective function (14a), the URLLC constraint (14b), and the combinatorial constraint (14d).

We consider a second objective of max-min rate fairness for all AGVs subject to the URLLC constraint as follows:
P1B:

$$\max_{\{\mathcal{B}_k\},\{\mathbf{w}_{k,j}\}} \min_{k\in\mathcal{K}} R_k(\mathbf{w}_{k,j}[t], \mathbf{h}_{k,j}[t]) \tag{15a}$$

$$\text{subject to: } \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k),\ \forall k \in \mathcal{K}, \tag{15b}$$

$$\sum_{k\in\mathcal{K}_j} \|\mathbf{w}_{k,j}[t]\|^2 \leq P_j,\ \forall j \in \mathcal{B}, \tag{15c}$$

$$\mathcal{B}_k[t] \subset \mathcal{B},\ \forall k \in \mathcal{K}. \tag{15d}$$

The max-min rate fairness in **P1B** improves the performance of the worst AGVs at the cost of total URLLC rate degradation, while the sum-rate maximization in **P1A** optimizes the total URLLC rate. Although max-min rate optimization can guarantee some fairness for users, it may not maximize the rate for all users. In practice, sum-rate maximization can be applied to the use-cases that require a high data rate for all users, while max-min rate optimization is rather applicable in the applications where users do not have minimum rate requirements. In other words, the max-min rate scheme is appropriate to reduce congestion in heavy traffic applications. For example, the AGVs have to perform compute-intensive tasks while suffering heavy traffic which can cause congestion in the network. In general, these two objective functions define the trade-offs between the network throughput and fairness. Therefore, they can be selected by the administrator depending on the use-case applications.

III. User-Centric CoMP Clustering

A. Problem Transformation

The beamformer variable of all gNBs that transmit to AGV k in cluster $\mathcal{B}_k$ $\mathbf{W}_k = \{\mathbf{w}_{k,j}, j \in \mathcal{B}_k\}$ is a matrix of $[M \times |\mathcal{B}_k|]$ continuous complex variables. Therefore, it is challenging to design joint clustering and beamforming solutions for all AGVs because these solutions consist of multiple matrices of continuous complex variables. In this paper, we propose using the codebook technique [11], [12], [19] so that the DRL agents can learn the transmit power and beam direction from a codebook instead of learning all the beamformer matrices for all AGVs.

The beamformer vector $\mathbf{w}_{k,j}$ from gNB j to AGV k can be decomposed into two separate parts as follows [12]:

$$\mathbf{w}_{k,j}[t] = \sqrt{p_{k,j}[t]}\bar{\mathbf{w}}_{k,j}[t], \tag{16}$$

where $p_{k,j}[t] = \|\mathbf{w}_{k,j}[t]\|^2$ denotes the transmit power of gNB j to AGV k at time slot t that satisfies constraint (14c), and $\bar{\mathbf{w}}_{k,j}[t]$ represents the beam direction of the transmit beamformer $\mathbf{w}_{k,j}[t]$. The beam direction vector $\bar{\mathbf{w}}_{k,j}[t]$ represents the degree of angles of the transmit beams with values in the range of $[0, 2\pi)$.

We consider a codebook $\mathcal{C} = [\mathbf{c}_q] \in \mathbb{C}^{M\times Q_{\text{code}}}$ composed of Q_{code} code vector $\mathbf{c}_q \in \mathbb{C}^{M\times 1}$. Each column of $\mathcal{C}$ is a

code that specifies a beam direction. The element of the codebook matrix is designed as follows [12]

$$c_{m,q} = \frac{1}{\sqrt{M}} \exp\left(i \frac{2\pi}{\Phi} \left\lfloor \frac{m \bmod (q + \frac{Q_{\text{code}}}{2}, Q_{\text{code}})}{Q_{\text{code}}/\Phi} \right\rfloor \right), \tag{17}$$

where $c_{m,q}$ refers to the phase shift of the nth antenna element in the qth code, Φ denotes the number of available phase values for each antenna element, and $\lfloor . \rfloor$ and mod(.) represent the floor and modulo operations, respectively.

The problem **P1A** in (14) can be rewritten as follows:

P2A:

$$\max_{\{\mathcal{B}_k\},\{p_{k,j}\},\{\bar{\mathbf{w}}_{k,j}\}} \sum_{k\in\mathcal{K}} R_k(\mathbf{w}_{k,j}[t], \mathbf{h}_{k,j}[t]) \tag{18a}$$

$$\text{subject to: } \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k),\ \forall k \in \mathcal{K}, \tag{18b}$$

$$\sum_{k\in\mathcal{K}_j} p_{k,j}[t] \leq P_j, \forall j \in \mathcal{B}, \tag{18c}$$

$$\mathcal{B}_k[t] \subset \mathcal{B}, \forall k \in \mathcal{K}, \tag{18d}$$

$$\bar{\mathbf{w}}_{k,j}[t] \in \mathcal{C}, \forall k \in \mathcal{K}, \forall j \in \mathcal{B}_k[t], \tag{18e}$$

The problem **P2A** is still difficult to solve because this kind of problem is NP-hard. To obtain the solution of problem **P2A**, we first decompose problem **P2A** into two subproblems, i.e., the CoMP clustering subproblem and beamforming design subproblem. In the following subsection, we propose a user-centric CoMP clustering algorithm to obtain the solution for the CoMP clustering subproblem.

B. User-Centric CoMP Clustering Algorithm

We design a user-centric clustering algorithm where each AGV k is served by a cluster of $\mathcal{B}_k$ gNBs. The cluster is defined on the reference signal's received power (RSRP) from gNBs. Adding gNBs to an existing cluster will increase the capacity of the cluster at the cost of additional complexity and signaling overhead. Therefore, it is important to balance CoMP efficiency and complexity. To minimize the signaling overhead in the fronthaul and CoMP server, we want to minimize the cluster size i.e., the number of coordinated gNBs per each cluster, while satisfying the SINR and URLLC constraints of each AGV. We determine a maximum number of gNBs in a cluster, i.e., $B_k \leq B_{\max}$ to balance the complexity against the CoMP efficiency trade-off.

Given the power allocation and codebook selection, the downlink SIR-protection level between the gNB j and AGV k is calculated as follows [13]

$$\rho_{k,j} = \frac{\mathbb{E}\left[\sum_{i\in\mathcal{J}_{k,j+}} \left|\mathbf{w}_{k,i}\mathbf{h}_{k,i}^H\right|^2\right]}{\mathbb{E}\left[\left|\mathbf{w}_{k,j}\mathbf{h}_{k,j}^H\right|^2 + \sum_{l\in\mathcal{J}_{k,j-}} \left|\mathbf{w}_{k,l}\mathbf{h}_{k,l}^H\right|^2\right]}, \tag{19}$$

where $\mathcal{J}_{k,j+}$ and $\mathcal{J}_{k,j-}$ denote the set of gNBs having higher and lower values of $\mathbb{E}\left[\left|\mathbf{w}_{k,i}\mathbf{h}_{k,i}^H\right|^2\right], i \neq j$ than the gNB j, respectively.

The user-centric CoMP clustering algorithm is presented in Algorithm 1. Algorithm 1 works as a greedy algorithm in which each AGV greedily searches all gNBs that satisfy the criteria (line 8-11).

Algorithm 1 User-centric CoMP Clustering

1: **Input**: $\{\mathbf{w}_{k,j}\}$, $\{\mathbf{h}_{k,j}\}$ $\forall k \in \mathcal{K}$ and $\forall j \in \mathcal{B}$, $B_{\max}$
2: Initialize $\mathcal{B}_k = \emptyset$ $\forall k \in \mathcal{K}$, $\rho_{\max}$
3: **for** AGV $k \in \mathcal{K}$ **do**
4: **for** gNB $j \in \mathcal{B}$ **do**
5: Generate the sorted list $\mathcal{J}_{k,j+}$ and $\mathcal{J}_{k,j-}$ for each pair (k, j)
6: Compute the downlink SIR-protection level $\rho_{k,j}$ for each pair (k, j)
7: **end for**
8: Sort the list $\{\rho_{k,j}\}$ in the descend order.
9: **for** $j \in \mathcal{B}$ **do**
10: $\mathcal{B}_k \leftarrow j$ if $\rho_{k,j} \leq \rho_{\max}$ and $B_k \leq B_{\max}$
11: **end for**
12: **end for**
13: **Output**: $\{\mathcal{B}_k\}$ $\forall k \in \mathcal{K}$

IV. Proximal Policy Optimization

In this section, we propose a DRL-based framework to obtain the solution for the beamforming design subproblem by modeling the beamforming design subproblem as a Markov Decision Process (MDP).

A. System State, Action, and Reward Design

Consider an infinite-horizon discounted MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathbf{Pr}, r, \gamma)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathbf{Pr} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability $r : \mathcal{S} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. The MDP of beamforming design can be characterized as follows:

1) The network state at time t is defined by the tuple $\mathcal{S} = (\{\mathcal{B}_k[t-1]\}_{k\in\mathcal{K}}, \{\mathbf{h}_{k,j}[t]\}_{k\in\mathcal{K},j\in\mathcal{B}})$ in which:
   - $\mathcal{B}_k[t-1], \forall k \in \mathcal{K}$ is the CoMP clustering at time $t-1$.
   - $\mathbf{h}_{k,j}[t], \forall k \in \mathcal{K}, \forall j \in \mathcal{B}$ is the channel state of all AGVs.
2) The action space at time t is the variables of problem **P2A** and defined by the tuple $\mathcal{A} = (\{p_{k,j}\}_{k\in\mathcal{K},j\in\mathcal{B}}, \{\bar{\mathbf{w}}_{k,j}\}_{k\in\mathcal{K},j\in\mathcal{B}})$. At each time t, the agent makes a decision of transmit power level and the corresponding codeword from gBNs to AGVs.
3) The reward is the signal from the environment to tell the agent how good the action is when it is executed. In the formulated problem, we aim to maximize the URLLC rate of the AGVs at each time slot. Naturally, the agent should take the transmission rates as its reward. However, if each gNB tries to maximize its transmission rate by increasing its transmit power, it can generate significant interference to other AGVs served by the other gNBs, hence, cannot satisfy the

URLLC constraint (18b) for all AGVs. Therefore, the reward for each AGV at time t is designed as follows:

$$r_k[t] = \begin{cases} \kappa_1 R_k[t] - \kappa_2 \sum_{j \in \mathcal{B}_k} P_{k,j}, \\ \kappa_3, \text{ if (18b) does not satisfy} \end{cases} \quad (20)$$

where the second term is the penalty for the gNBs for the exceeding transmit power in the cluster $\mathcal{B}_k$ and κ_1 and κ_2 are tunable scale coefficients. Moreover, κ_3 is a negative reward to penalize the agent if the URLLC constraint in (18b) cannot be satisfied. Reward for each agent j (i.e., gNB j) is formulated as follows:

a) Maximize sum-rate: Reward for each agent j (i.e., gNB j) is the sum of reward of all the AGVs served by gNB j:

$$r_j[t] = \sum_{k \in \mathcal{K}_j} r_k[t]. \quad (21)$$

b) Maximize minimum rate: Reward for each agent j (i.e., gNB j) is the minimum reward of the AGVs served by gNB j:

$$r_j[t] = \min_{k \in \mathcal{K}_j} r_k[t]. \quad (22)$$

B. Proximal Policy Optimization

Proximal policy optimization (PPO) [20] is a model-free, online, on-policy, policy gradient reinforcement learning method. This algorithm is a type of policy gradient training that alternates between sampling data through environmental interaction and optimizing a clipped surrogate objective function using stochastic gradient descent (SGD).

PPO alternatively constructs an unconstrained surrogate objective function to remove the incentive for large policy updates. PPO updates policies by taking multiple steps of (usually minibatch) SGD to maximize the objective

$$\theta^{(n+1)} = \arg\max_{\theta} \mathop{\mathrm{E}}_{s,a \sim \pi_{\theta_n}} \left[L(s, a, \theta_k, \theta) \right], \quad (23)$$

where L is given in (24). $\pi_\theta(a|s)$ is new parameterized policy trying to seek the optimal parameter vector θ, and $\pi_{\theta_n}(a|s)$ is the old policy. Here, ϵ is a small hyperparameter presenting how far the new policy is allowed to go from the old policy. The advantage function $A^{\pi_{\theta_n}}(s, a)$ can be calculated by

$$A^{\pi_{\theta_n}}(s, a) = Q^{\pi_{\theta_n}}(s, a) - V^{\pi_{\theta_n}}(s), \quad (27)$$

where $Q^{\pi_{\theta_n}}(s, a)$ is the action-value function estimated by samples, and $V^{\pi_{\theta_k}}(s)$ is the approximation of the state-value function.

$$Q^{\pi_{\theta_n}}(s_t, a_t) = \mathbb{E}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l}) \right]. \quad (28)$$

The PPO algorithm is presented in Algorithm 2 and illustrated in Fig. 3. Each agent (i.e., gNB) collects a minibatch of transitions by running the current policy to produce the beamforming actions including the transmit power and beam direction to its connected AGVs (line 4). Each agent computes advantage estimates and updates the policy by maximizing the PPO-clip objective with the minibatch of transitions (line 5,6). Then, each agent trains the value functions by regression on mean-squared error (MSE) (line 7). The steps are repeated until the agents' policies converge to stationary policies.

Algorithm 2 PPO-based beamforming design

1: Initialize policy parameter $\theta^{(0)}$, initialize value function parameters $\phi^{(0)}$ for each gNB agent;
2: **for** $n = 0, 1, 2, ..$ iterations **do**
3: **for** each gNG agent **do**
4: Collect a minibatch of D transitions $\mathcal{D}_n = \{s_i, a_i, r_i, s_{i+1}\}_{i=0:D-1}$ by running policy π_θ;
5: Compute advantage estimates $\hat{A}(s_t, a_t)$ based on the current value function $V_{\phi^{(n)}}(s_t)$;
6: Update the policy by maximizing the PPO-clip objective in (25) where

$$g(\epsilon, A) = \begin{cases} (1+\epsilon)A & A \geq 0 \\ (1-\epsilon)A & A < 0. \end{cases} \quad (29)$$

7: Fit value function by regression on MSE in (26)
8: **end for**
9: **end for**

V. Game Theoretic-Based CoMP Clustering

The user-centric CoMP clustering Algorithm 1 in Section III is based on the downlink SIR-protection criteria which is the maximum average SIR that an AGV can potentially achieve by removing the number of gNBs with lower signal strength from the original cluster [13]. However, this criteria does not consider the interference that affects other AGVs. This means that the current user-centric clustering mechanism is selfish and does not guarantee an equilibrium for all AGVs. The game theoretic-based solution can obtain equilibria for all AGVs.

There is an increasing interest in applying game theory to design self-organized, distributed cooperative clustering [21], [22]. In the game-theoretic approach, a payoff function is introduced to formulate the CoMP gain and cost trade-off for forming CoMP clusters.

A. Clustering Game Formulation

We consider a CoMP clustering game in which each AGV is a player trying to select a set of serving gNBs to maximize its payoff. We define the action of each player AGV as follow:

$$a_{k,j} = \begin{cases} 1, \text{ if AGV } k \text{ selects gNB } j \\ 0, \text{ otherwise.} \end{cases}$$

The clustering game can be formulated as follows:

Definition 1. *The CoMP clustering game is a tuple* $\mathcal{G} = (\mathcal{K}, \{a_{k,j}\}, \{\mathcal{P}_{k,j}\})$ *where*

1) *Player set: set of AGV* $\mathcal{K}$.
2) *Strategy: the strategy of each player is defined as decisions on choosing a set of gNBs to be served* $\boldsymbol{A} =$

$$L(s,a,\theta_k,\theta)=\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_n}(a|s)}A^{\pi_{\theta_n}}(s,a),\ \ \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_n}(a|s)},1-\epsilon,1+\epsilon\right)A^{\pi_{\theta_n}}(s,a)\right), \tag{24}$$

$$\theta^{(n+1)}=\arg\max_{\theta}\frac{1}{|\mathcal{D}_n|\Delta T}\sum_{\tau\in\mathcal{D}_n}\sum_{t=0}^{\Delta T}\min\left(\frac{\pi_\theta(t|s_t)}{\pi_{\theta_k}(a_t|s_t)}A^{\pi_{\theta_k}}(s_t,a_t),g(\epsilon,A^{\pi_{\theta_k}}(s_t,a_t))\right); \tag{25}$$

$$\boldsymbol{\phi}^{(n+1)}=\arg\min_{\phi}\frac{1}{|\mathcal{D}_n|\Delta T}\sum_{\tau\in\mathcal{D}_n}\sum_{t=0}^{\Delta T}\left[V_{\boldsymbol{\phi}^{(n)}}(\boldsymbol{s}[t])-r(\boldsymbol{s}[t],\boldsymbol{a}[t])\right]^2; \tag{26}$$

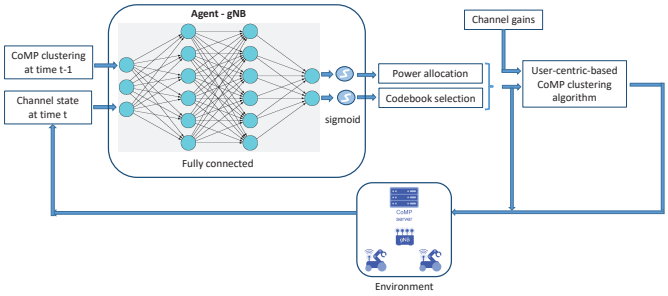


Figure 3: Joint CoMP clustering and beamforming design framework.

$\{\boldsymbol{a}_j\}_{j\in\mathcal{B}}$, $\boldsymbol{a}_j=\{a_{k,j}\}_{k\in\mathcal{K},j\in\mathcal{B}}$ *to maximize its payoff function.*

3) *Payoff function: The payoff of player k is given by*

$$\mathcal{P}_k(\boldsymbol{A})=\sum_{j\in\mathcal{B}}\mathcal{P}_{k,j}(\boldsymbol{a}_j), \tag{30}$$

$$\mathcal{P}_{k,j}(\boldsymbol{a}_j)=\frac{\left(a_{k,j}\left|\boldsymbol{w}_{k,j}\boldsymbol{h}_{k,j}^H\right|^2\right)^\alpha}{\sum_{l\in\mathcal{K}}\left(a_{l,j}\left|\boldsymbol{w}_{l,j}\boldsymbol{h}_{l,j}^H\right|^2\right)^\alpha}-\xi a_{k,j}\sum_{l\neq k\in\mathcal{K}}\left|\boldsymbol{w}_{k,j}\boldsymbol{h}_{l,j}^H\right|^2, \tag{31}$$

where α and ξ are positive. The first term of the payoff function presents the percentage allocated power of gNB j to AGV k. The second term presents the total interference caused by the transmission from gNB j to AGV k. The payoff function indicates that the utility and the total interference each AGV incurs will vary inversely according to the increasing number of AGVs connected to the same gNB.

In the following subsections, we transform the game $\mathcal{G}$ into a mean-field game and analyze the Nash equilibrium.

B. Mean Field Approximation for CoMP Clustering

When the system becomes large, traditional game-theoretic analysis is computationally inefficient because every single action of every player should be taken into account. A mean-field game is proposed to tackle the dimensionality difficulty of the traditional game by taking the statistical mean-field distribution instead of tracking the action of each player [23].

Denote the weight by $\omega_{k,j}=|\mathbf{w}_{k,j}\mathbf{h}_{k,j}^H|^2$, we define the mean-field as a weighted α-norm of all the actions as follows:

$$m_j=\left(\frac{1}{K}\sum_{k\in\mathcal{K}}\left(\omega_{k,j}a_{k,j}\right)^\alpha\right)^{\frac{1}{\alpha}},\forall j\in\mathcal{B}. \tag{32}$$

The payoff function in (31) can be rewritten as follows:

$$\begin{aligned}\mathcal{P}_{k,j}(a_{k,j},m_{j,-k})&=\frac{1}{K}\left(\frac{\omega_{k,j}a_{k,j}}{m_j}\right)^\alpha-\xi\mathcal{I}_{k,j}a_{k,j},\\&=\frac{\left(\omega_{k,j}a_{k,j}\right)^\alpha}{(K-1)m_{j,-k}^\alpha+\left(\omega_{k,j}a_{k,j}\right)^\alpha}-\xi\mathcal{I}_{k,j}a_{k,j},\end{aligned} \tag{33}$$

where $\mathcal{I}_{k,j}=\sum_{l\neq k\in\mathcal{K}}\left|\mathbf{w}_{k,j}\mathbf{h}_{l,j}^H\right|^2$, and

$$\begin{aligned}m_{j,-k}^\alpha&=\frac{1}{K-1}\sum_{j\neq k}\left(\omega_{l,j}a_{l,j}\right)^\alpha\\&=\frac{K}{K-1}\left(m_j^\alpha-\frac{\left(\omega_{k,j}a_{k,j}\right)^\alpha}{K}\right).\end{aligned} \tag{34}$$

The payoff function of a player has the following properties:

- The payoff function depends only on the player's action $a_{k,j}$ and the mean field m_j.
- The payoff is discontinuous when there is no connection to the gNB j, i.e., $\sum_{k\in\mathcal{K}}\left(\omega_{k,j}a_{k,j}\right)^\alpha=0$.

C. Equilibrium for Clustering Game

In this section, we characterize the mean-field equilibrium of the formulated game.

Definition 2. *An action vector $\boldsymbol{a}_j^{NE}=\{a_{k,j}^{NE}\}_{k\in\mathcal{K}}$ is said to be a Nash equilibrium if no player can improve its payoff by unilaterally deviating its action from the Nash equilibrium, such that:*

$$\mathcal{P}_{k,j}(a_{k,j}^{NE},m_{j,-k})\geq\mathcal{P}_{k,j}(a_{k,j},m_{j,-k}),\ a_{k,j}\in(0,1),\forall k.$$

Theorem 1. *There exists at least one Nash equilibrium for the game $\mathcal{G}$.*

Proof. We consider the case there is at least one AGV connected to a gNB so that the payoff function is smooth, continuous and differential. If there is no AGV in the coverage of an gNB, the game $\mathcal{G}$ simply excludes such gNB out of the strategy of the players, i.e., AGVs.

The first and second derivative with respect to $a_{k,j}$ can be written as follows:

$$\begin{aligned}\frac{\partial \mathcal{P}_{k,j}(a_{k,j}, m_j)}{\partial a_{k,j}} &= \frac{1}{K}\left[\frac{\alpha\omega_{k,j}^{\alpha} a_{k,j}^{\alpha-1} m_j^{\alpha} - \left(\omega_{k,j} a_{k,j}\right)^{\alpha}\left(\frac{\alpha}{K} a_{k,j}^{\alpha-1}\right)}{m_j^{2\alpha}}\right] - \xi\mathcal{I}_{k,j} \\ &= \frac{\alpha\omega_{k,j}^{\alpha} a_{k,j}^{\alpha-1}}{K}\left[\frac{m_j^{\alpha} - \frac{a_{k,j}^{\alpha}}{K}}{m_j^{2\alpha}}\right] - \xi\mathcal{I}_{k,j},\end{aligned} \tag{35}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{P}_{k,j}(a_{k,j}, m_j)}{\partial {a_{k,j}}^2} &= \frac{\alpha\omega_{k,j}^{\alpha}}{K}\left(m_j^{\alpha} - \frac{a_{k,j}^{\alpha}}{K}\right) \\ &\quad \times \left[\frac{(\alpha-1)a_{k,j}^{\alpha-2} m_j^{2\alpha} - a_{k,j}^{\alpha-1}\frac{2\alpha}{K} a_{k,j}^{\alpha-1} m_j^{\alpha}}{m_j^{4\alpha}}\right] \\ &= \frac{\alpha\omega_{k,j}^{\alpha}}{K}\left(m_j^{\alpha} - \frac{a_{k,j}^{\alpha}}{K}\right)\frac{a_{k,j}^{\alpha-2}}{m_j^{3\alpha}}\left[(\alpha-1)m_j^{\alpha} - \frac{2\alpha}{K} a_{k,j}^{\alpha}\right].\end{aligned} \tag{36}$$

For $0 \leq \alpha \leq 1$, the second derivative of the payoff function with respect to $a_{k,j}$ is negative, then the payoff is concave with respect to own-action $a_{k,j}$. Therefore, there exists at least one Nash equilibrium for the game $\mathcal{G}$. □

We consider an asymmetric game in which all the players have asymmetric strategies in equilibrium whenever it exists. In other words, each AGV has its own clustering strategy which is different from other AGVs. However, due to the complicated structure of the payoff function, deriving a closed-form asymmetric Nash equilibrium is not trivial. Instead, we propose an iterative form that converges to mean-field equilibrium. As the number of players tends to infinity, the mean-field equilibrium asymptotically converges to Nash equilibrium.

Definition 3. *Mean field best response of player k given the actions of other players given by*

$$\boldsymbol{Br}(a_{k,j}, m_j) = \arg\max_{a_{k,j}} \left[\frac{1}{K}\left(\frac{\omega_{k,j} a_{k,j}}{m_j}\right)^{\alpha} - \xi\mathcal{I}_{k,j} a_{k,j}\right]. \tag{37}$$

Theorem 2. *The iterative best response updates converge to Nash equilibrium*

$$a_{k,j}(\tau+1) = \lambda(\tau)\boldsymbol{Br}(a_{k,j}(\tau), m_j(\tau)) + (1-\lambda(\tau))a_{k,j}(\tau), \tag{38}$$

where τ represents the iterations and $\lambda(\tau)$ is a step size and

$$\boldsymbol{Br}(a_{k,j}(\tau), m_j(\tau)) = \left[K\left(m_j^{\alpha}(\tau) - \frac{K\xi\mathcal{I}_{k,j} m_j^{2\alpha}(\tau)}{\alpha\omega_{k,j}^{\alpha} a_{k,j}^{\alpha-1}(\tau)}\right)\right]^{\frac{1}{\alpha}}, \tag{39}$$

$$\begin{aligned}m_j(\tau+1) = \lambda(\tau)&\left[\frac{a_{k,j}^{\alpha}(\tau)}{K} + \frac{K\xi\mathcal{I}_{k,j} m_j^{2\alpha}(\tau)}{\alpha\omega_{k,j}^{\alpha} a_{k,j}^{\alpha-1}(\tau)}\right]^{\frac{1}{\alpha}} \\ &+ (1-\lambda(\tau))m_j(\tau).\end{aligned} \tag{40}$$

Proof. Since $\mathbf{Br}(a_{k,j}(\tau), m_j(\tau))$ is obtained by setting the first derivative of the payoff function in (35) equals zero, then it is the unique solution.

Algorithm 3 Distributed Game-based CoMP Clustering

1: Initialize $a_{k,j}(0)$ and $m_j(0)$ $\forall k \in \mathcal{K}, j \in \mathcal{B}$;
2: All gNBs broadcast their beamforming profiles $\mathbf{w}_{k,j}$;
3: **repeat**
4: Each AGV k updates its strategy $a_{k,j}(\tau)$ according to (38) and (39);
5: Update mean field according to (40);
6: **until** $|a_{k,j}(\tau+1) - a_{k,j}(\tau)| \leq \varepsilon$

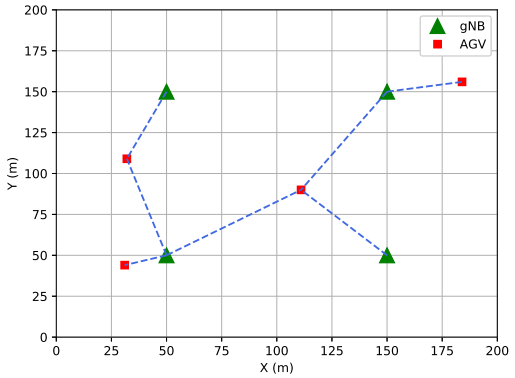


Figure 4: CoMP clustering illustration. The dash lines present the association between the gNBs and the AGVs. Each AGV has its own CoMP cluster indicated by the set of associated gNBs.

The iterative best response update (38) has the form of Ishikawa (Mann) iteration [24]. It was proven in [24] that, with a vanishing learning rate, i.e., $\lambda(\tau) > 0$, $\sum_{\tau}\lambda(\tau) = \infty$, and $\sum_{\tau}\lambda^2(\tau) < \infty$, the iterative best response update (38) converges strongly to a fixed point which is a unique Nash equilibrium. □

A distributed game-based CoMP clustering is presented in Algorithm 3. After receiving the beamforming information from gNBs, each AGV updates its strategy by the iterative best response equation and the approximated mean-field value without knowledge of other AGVs' actions. Therefore, this method can reduce the message exchange overhead and complexity of the algorithm.

D. Complexity

In practice, PPO usually is implemented in Actor-Critic framework in which the policy network is implemented as an actor and the value function is implemented as a critic network. The computational complexity of the PPO-based algorithm can be calculated based on the complexity related to the training of the actor and critic neural networks. Let L^{actor} and L^{critic} denote the number of fully connected layers of the actor network and critic network, respectively. The computational complexity of the PPO-based algorithm is $\mathcal{O}(\sum_{l=0}^{L^{\text{actor}}-1} u_l^{\text{actor}} u_{l+1}^{\text{actor}} + \sum_{l=0}^{L^{\text{critic}}-1} u_l^{\text{critic}} u_{l+1}^{\text{critic}})$ [25], where u_l^{actor} and u_l^{critic} are the unit numbers in the l-th layers of actor network and in the l-th layer of critic network, respectively. Here, u_0^{actor} and u_0^{critic} represent the

Table I: PPO hyperparameters setting

Parameter	Value
Policy network	128, relu, 128, relu, 128, relu, tanh
Value network	128, relu, 128, relu, 128, relu, linear
Step size	$1e-3$
Batch size	20
Discount factor	0.995
Epsilon clip	0.1

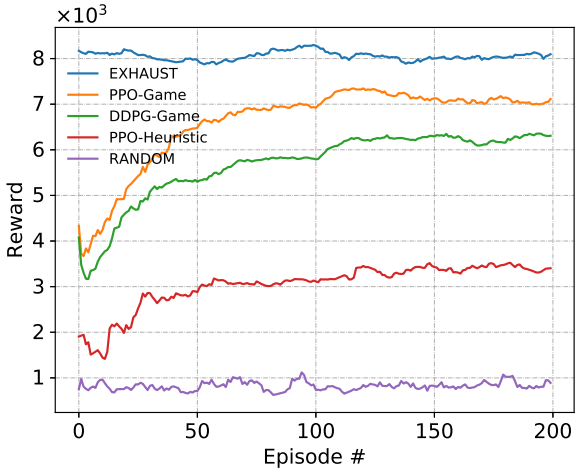


Figure 5: Accumulative reward.

input sizes of actor network and critic network, respectively. The input size of the actor and critic networks in our model is $u_0^{\text{actor}} = u_0^{\text{critic}} = |\mathcal{B}| + M \times |\mathcal{B}| \times |\mathcal{K}|$ where M is the number of antennas of each gNB, $|\mathcal{K}|$ is the number of AGVs, and $|\mathcal{B}|$ is the number of gNBs. It can be seen that, the computation complexity of the PPO-based algorithm increases according to the network state, i.e., the number of AGVs and the number of gNBs.

The computational complexity of Algorithm 3 is a polynomial function of the number of iterations of the iterative best response update (38), i.e., $\mathcal{O}(2T \times |\mathcal{K}| \times |\mathcal{B}|)$ where T is the number of iterations. It is proved in Theorem 2 that T is finite and in the simulation, we see that the number of iterations T of the best response update (38) is around 5.

VI. Simulation Results

A. Simulation Setting

We perform extensive simulations to evaluate the performance of our proposed design in terms of the sum URLLC rate in (9), i.e., the objective of the optimization problem **P1**. We vary the number of AGVs in the range of [2-20] in a 200×200 meters square automated warehouse. There are 4 gNBs each with 4 antennas so that they can fully cover the area and provide service to the AGVs as depicted in Fig. 4. At the beginning of each episode, the central controller generates a uniformly distributed destination for each AGV and the AGV follows the shortest path from its starting point to its destination. The velocity of each AGV follows a Gaussian distribution $\mathcal{N}(5, 2)$ with a mean 5 m/s and a standard deviation of 2. The carrier frequency

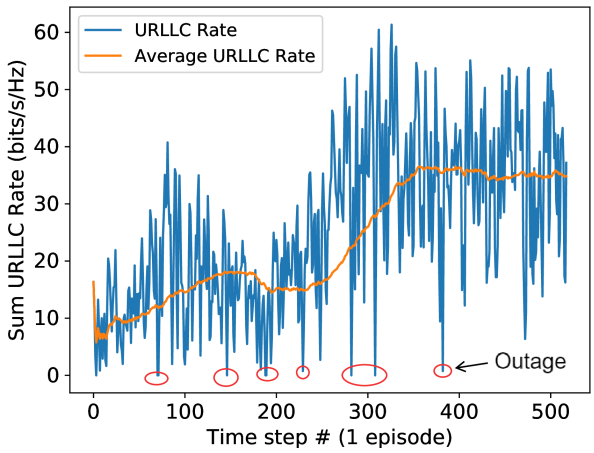


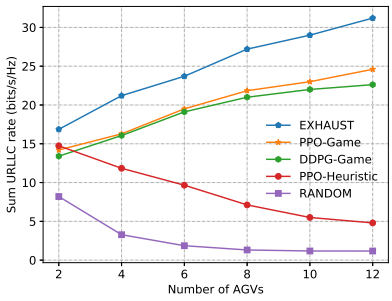
Figure 6: URLLC rate performance during one episode with around 500 time steps

is 6 GHz with 2 MHz bandwidth. The pathloss exponent is set to 3.76, the noise power spectral density is set to -174 dBm/Hz and the decoding error probability is set to 10^{-9}. The data packet size is 20 bytes and channel blocklength is 512 symbols [26].

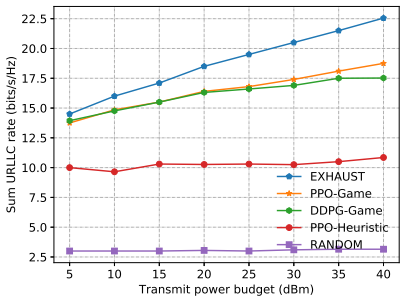
To implement the neural networks, we employ the powerful open-source machine learning framework PyTorch version 1.2.0 primarily developed by Facebook's AI Research lab. The programming language is Python 3.7 on a desktop computer with hardware configuration: Intel(R) Core(TM) i7-4790 CPU 3.60GHz and 16GB RAM. The hyperparameters of our proposed PPO Algorithm 2 is presented in Table I. However they are not chosen arbitrarily but should be related to the network parameters. For example, we choose the number of AGVs in the range of [2-20]. Therefore, the number of hidden layers is relatively small and is tuned in the range of 1 to 3, and the number of hidden units is from 64 to 128.

We compare our proposed joint CoMP clustering and beamforming design scheme (denoted as 'PPO-Game') with four benchmark schemes as follows:

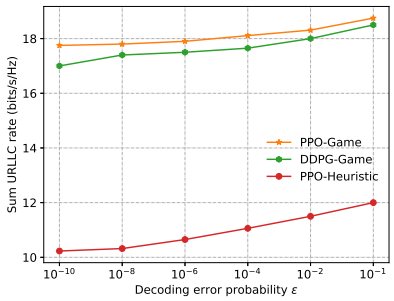
- 'DDPG-Game': This baseline is the multi-agent off-policy deep deterministic policy gradient [25], [27] combined with the distributed game theoretic based CoMP clustering Algorithm 3. We investigate whether on-policy or off-policy gradient method outperforms in a dynamic environment as in a robotic network.
- 'PPO-Heuristic': The user-centric CoMP clustering in the Algorithm 1 works as a greedy algorithm where each AGV selfishly searches a set of serving gNBs satisfies downlink protection criteria and without considering interference caused to other AGVs.
- 'EXHAUST': We use the exhaustive search method over the Euclidean space $\mathcal{K} \times \mathcal{B} \times \mathcal{C} \times P$. The 'EXHAUST' baseline is considered the optimal solution for the formulated problem.
- 'RANDOM': The CoMP clustering, transmit power, and beam direction (codebook) are randomly selected



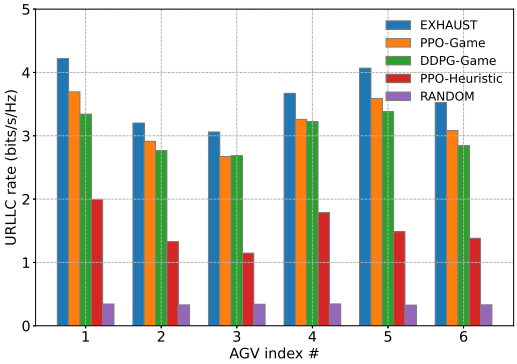
(a) vs number of AGVs.



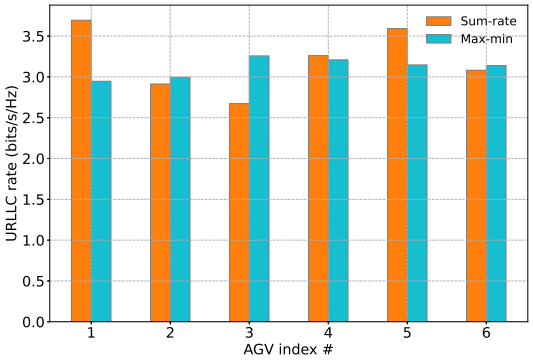
(b) vs transmit power budget



(c) vs decoding error probability ϵ_k



(d) URLLC rate of each AGV (6 AGVs)



(e) Sum-rate vs Max-min rate (6 AGVs)

Figure 7: URLLC rate performance.

at each time slot.

The performance of the proposed scheme and the baseline is evaluated through the following metrics:

1) URLLC rate: This metric is the URLLC rate in (9).
2) Outage probability: This metric is the percentage of the solutions that do not satisfy the URLLC constraint (14b) and (18b).
3) Complexity: This metric is the computation complexity of each baseline, and signaling overhead.

Note that, except Fig. 7(e), in all other Figures, the 'PPO-Game' scheme is simulated with the 'Sum-rate' objective.

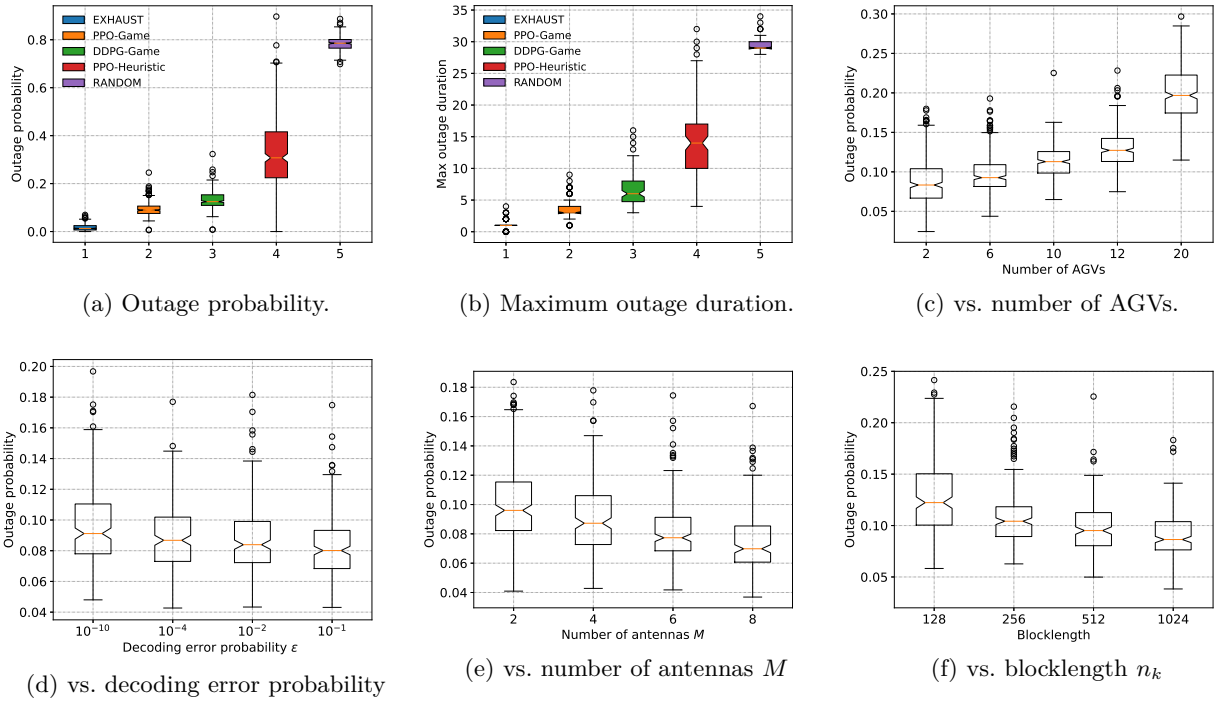
B. Results Analysis

Fig. 4 illustrates the CoMP cluster of each AGV created by our proposed 'PPO-Game' scheme. The set of gNBs associated with an AGV forms a CoMP cluster of that AGV. It can be observed that the CoMP cluster of each AGV is different from the others depending upon the channel condition and beamforming of each gNB.

1) Convergence Performance: Fig. 5 shows the convergence of the accumulative reward of our proposed scheme and four benchmark schemes over 200 episodes (each with hundreds of time steps). The 'EXHAUST' scheme achieves the highest reward while the 'RANDOM' experiences the worst performance. Our proposed scheme 'PPO-Game' improves gradually over the episodes and converges to a fairly stable situation in approximately 150 episodes. It can be observed that our proposed scheme 'PPO-Game' significantly outperforms the 'PPO-Heuristic' baseline and reaches a stable reward close to the 'EXHAUST' baseline which is the optimum. Moreover, the 'DDPG-Game' baseline has a similar convergence behavior but converges to a lower value compared to the 'PPO-Game' scheme. In a stable environment, the off-policy DDPG-based algorithm may outperform the on-policy PPO-based algorithm due to the sample efficiency characteristic of the DDPG method. However, in a highly dynamic environment, which is the case in this paper, the DDPG method may cause sudden failures due to the exploration noise, resulting in instabilities during training due to the sensitivity to the model hyperparameters [28]. Whilst the on-policy PPO method monotonically improves the policy and guarantees the new policy after the gradient step is not too different than before [29].

Fig. 6 shows the sum URLLC rate within an episode of around 500 time steps. We can observe that there are some time steps at which the outage happens, i.e., the AGVs do not satisfy the URLLC constraint (14b) and (18b) in problem **P1A** and **P2A**. Such outage happens when the AVGs fail to obtain the Nash equilibrium for the CoMP clustering game, or the PPO agents produce a beamforming profile that does not satisfy the URLLC constraint. For example, when an AGV is at the edge of a gNB's coverage but not in any other gNBs' coverage. The outage performance is analyzed in detail with different scenarios in Fig 8.

This article has been accepted for publication in IEEE Transactions on Network and Service Management. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2022.3202168



(a) Outage probability. (b) Maximum outage duration. (c) vs. number of AGVs.

(d) vs. decoding error probability (e) vs. number of antennas M (f) vs. blocklength n_k

Figure 8: Outage probability performance.

2) URLLC Rate Performance: Fig. 7(a) depicts the sum URLLC rates of the five schemes versus the number of AGVs. The proposed scheme 'PPO-Game' baseline can handle the interference caused by the overlapping clusters. Therefore, when the number of AGVs increases, the sum URLLC rate of all AGVs increases. A similar increasing trend can be seen with the 'DDPG-Game' baseline. This result implies an effective adaptation of the game-theoretic CoMP clustering according to the increasing number of AGVs.

Moreover, we can see that when the number of AGVs in the network is small, the performance gap between PPO and DDPG methods is relatively small. However, when the number of AGVs becomes larger, the performance gap between these two policy gradient methods is more significant. When the number of AGVs is 12, the 'PPO-Game' achieves a sum URLLC rate of 24.65 bits/s/Hz compared to 31.2 bits/s/Hz of the 'EXHAUST' baseline, in other words, a performance of approximately 78.5% compared to the optimal solution. On the other hand, the 'PPO-Heuristic' baseline experiences a noticeable decrease of sum URLLC rate when the number of AGVs increases. The poor performance of the 'PPO-Heuristic' baseline can be explained by the fact that the interference is not managed in this clustering algorithm, even though both 'PPO-Game' and 'PPO-Heuristic' schemes implement the same PPO-based beamforming algorithm. The user-centric CoMP clustering in the 'PPO-Heuristic' baseline works as a greedy algorithm in which each AGV greedily searches all the possible serving gNBs that satisfy the SIR criteria while ignoring the interference that may cause to the other AGVs. Therefore, the more AGVs in the network, the more interference each AGV incurs, and the poorer network is.

Fig. 7(b) plots the sum URLLC rates versus the transmit power budget. It can be seen that the sum URLLC rates of the 'PPO-Game' scheme, 'DDPG-Game' and 'EXHAUST' baseline increase along with the increase in the transmit power budget whereas the sum URLLC rate of the 'PPO-Heuristic' baseline is nearly constant. This result again confirms the 'PPO-Game' scheme can manage interference better than the 'PPO-Heuristic' baseline. When the transmit power increases the interference also increases, hence, an interference adaptive scheme would be beneficial.

Moreover, it can be observed that when the transmit power budget is small, the performances of 'PPO-Game' and 'DDPG-Game' schemes are almost identical. However, when the transmit power budget increases, the performance gap between 'PPO-Game' and 'DDPG-Game' schemes increase significantly.

In Fig. 7(c), we draw the sum URLLC rates of three schemes 'PPO-Game', 'DDPG-Game' and 'PPO-Heuristic' versus the decoding error probability $\epsilon = \epsilon_k, \forall k$. For all considered schemes, the sum URLLC rate is a monotonically increasing function of the decoding error probability. This is because the inverse error function $Q^{-1}(\epsilon)$ is a monotonically decreasing function of ϵ. However, as can be observed, the impact of the decoding error probability on the URLLC rate is minor. As we can see in (9), the second term can be interpreted as a penalty on the rate in order to guarantee the decoding error probability in a finite blocklength regime. This penalty is relatively small compared to the Shannon capacity, i.e., the first term in (9). Moreover, both 'PPO-

Game' and 'DDPG-Game' schemes significantly outperform the 'PPO-Heuristic' scheme, and the observed performance gain is similar to the simulation scenarios with the transmit power budget (7(b)) and the number of AGVs (7(a)).

Fig. 7(d) presents URLLC rate of each AGV (6 AGVs) in five schemes. It can be seen that the 'EXHAUST' scheme has the highest URLLC rate for all AGVs compared to all other schemes and the 'RANDOM' baseline has the lowest URLLC rate for all AGVs. The 'PPO-Game' scheme has a slightly higher URLLC rate than that of the 'DDPG-Game' scheme. A significant improvement for all AGVs obtained by the 'PPO-Game' and 'DDPG-Game' schemes compared to the 'PPO-Heuristic' scheme can be explained by the fact that all the AGVs can find the equilibrium in the CoMP clustering game, and the DRL agents can produce beamforming profiles that adapt to the changing of the network state of each AGV.

In Fig. 7(e), we compare the URLLC rate performance of our proposed 'PPO-Game' scheme with two different objective functions, i.e., maximize sum-rate (denoted as 'Sum-rate') and maximize minimum rate (denoted as 'Max-min'). It can be seen that with the 'Max-min' objective the proposed 'PPO-Game' scheme can achieve better fairness compared to the 'Sum-rate' objective. However, the 'Sum-rate' objective provides a higher total throughput (sum URLLC rate) of all the AGVs than the 'Max-min' objective. More specifically, the 'Max-min' objective achieves a better URLLC rate for the worst user, i.e., AGV number 3, than the 'Sum-rate' objective, but achieves a lower URLLC rate for AGVs number 1 and number 5. Note that, with the equilibrium obtained in the CoMP clustering game our proposed 'PPO-Game' scheme can achieve certain fairness compared to the 'EXHAUST' scheme as shown in Fig. 7(d).

3) Outage Probability Performance: Fig 8(a) shows the outage probability of all schemes with 5 AGVs for all episodes. As expected, the 'EXHAUST' baseline has the lowest outage probability at around the median value of 2% while the 'RANDOM' baseline has the highest outage probability at around 78%. The 'PPO-Game' scheme has a lower outage probability than that of the 'DDPG-Game' scheme, at around 9% and 13%, respectively. The 'PPO-Heuristic' baseline has the median value of outage probability at around 30% but it has the widest range of outage probability value compared to all other schemes, which is from 0% to 72% with some outliers over 90%. In the worst case, the 'PPO-Heuristic' baseline can obtain a poor performance as the 'RANDOM' baseline. This result once again confirms our proposed scheme 'PPO-Game' can obtain a comparable performance compared to the exhaustive search algorithm which can be considered an optimal solution.

Fig. 8(b) compare the maximum continuous outage duration of five schemes with 10 AGVs for all episodes. Similarly to the outage probability shown in Fig 8(a), the 'EXHAUST' scheme has the lowest outage duration at around the median value of 1 time step. In contrast, the 'RANDOM' baseline has the highest outage duration at around the median value of 29 time steps. The 'PPO-Game' scheme has the outage duration value close to the 'EXHAUST' scheme at around 3 time steps, while the outage duration of the 'DDPG-Game' and 'PPO-Heuristic' schemes are two times and five times higher than that of the 'PPO-Game' scheme, respectively. This result again confirms the performance gap between the 'PPO-Game' and 'DDPG-Game' schemes is more significant when the number of AGVs increases.

We investigate the performance in terms of outage probability of our proposed scheme 'PPO-Game' with the variation of the number of AGVs, decoding error probability ϵ, the number of antennas M, and blocklength n_k, in Fig. 8(c), Fig. 8(d), Fig. 8(e), and Fig. 8(f), respectively. The outage probability values are collected over 300 running episodes, each episode is with hundreds of time steps.

In Fig. 8(c), it can be observed that with the fixed wireless resource, i.e., system bandwidth, and a number of serving gNBs in the network, the more number of AGVs the more probability the AGVs incur outage. The outage probability increases gradually with a small number of AGVs but increases dramatically when the number of AGVs is sufficiently large. Therefore, to assure the URLLC can be achieved when a large number of AGVs operates in the network, it is important to guarantee enough wireless resources and a number of serving gNBs.

In Fig. 8(d), we plot the outage probability of our proposed scheme with the different values of the decoding error probability requirement. As we can see, a tighter reliability requirement and a higher outage probability will be obtained. However, the increase of the outage probability when we decrease the decoding error probability requirement is not significant. For example, when the decoding error probability requirement is 10^{-1}, the outage probability median value is 0.08 or 8% but when the decoding error probability requirement decreases to 10^{-10} the outage probability only increases to around 0.09 or 9%.

In Fig. 8(e), the more number of antennas in each gNBs, the lower outage probability we can achieve. This is because we can obtain a better SINR value with a higher number of antennas, therefore, a lower outage probability will be incurred. A similar trend can be observed in Fig. 8(f) when we increase the blocklength. A higher blocklength value, a lower outage probability we can obtain. However, we cannot increase more blocklength values to achieve a better outage probability performance since the achievable URLLC rate will be saturated when the blocklength value exceeds 1024.

4) Complexity Performance: In Fig. 9, we compare the complexity in terms of computation time in milliseconds of four schemes except for the 'RANDOM' baseline. It is obvious the 'EXHAUST' baseline has the highest complexity compared to all other schemes. The complexity of the 'EXHAUST' baseline increases significantly with the number of AGVs in the network, whereas the complexities of the 'PPO-Game', 'DDPG-Game', and 'PPO-Heuristic' schemes increase slightly. For example, when the number of AGVs is 2, the complexity of the 'EXHAUST' baseline is about 3 times higher than that of the 'PPO-Game', 'DDPG-Game', and 'PPO-Heuristic' schemes. However, when the

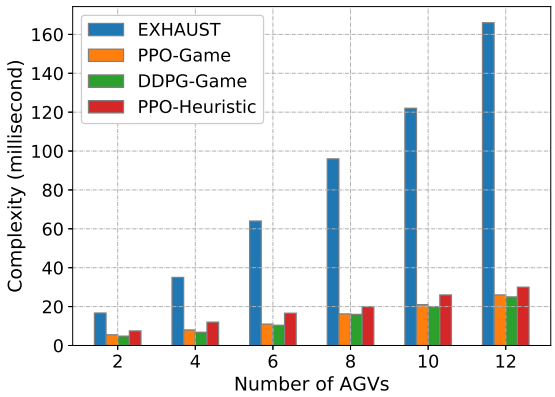


Figure 9: Complexity comparison

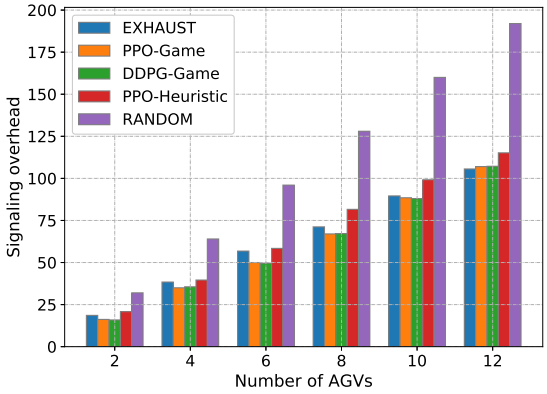


Figure 10: Signaling overhead

number of AGVs in the network is 12, the complexity of the 'EXHAUST' baseline is about 6 times higher than the other schemes. Furthermore, while the 'PPO-Game' and 'DDPG-Game' schemes have similar complexity, the 'PPO-Heuristic' scheme has a higher complexity than that of the 'PPO-Game' and 'DDPG-Game' schemes. This result can be explained by the fact that the PPO and DDPG algorithms are implemented by the actor-critic method, i.e., using two neural networks to implement the policy network and value network separately. Moreover, as stated in Section III.B and Section V.D, we see that the heuristic user-centric CoMP clustering algorithm (Alg.1) has a slightly higher complexity than that of the Game-based CoMP clustering algorithm (Alg.3.)

Fig. 10 depicts the signaling overhead of five schemes. The signaling overhead is calculated based on the number of connections between the AGVs and gNBs and the 5G-RRC (Radio Resource Control) connection setup procedure, i.e., 8 messages over a 5G-RRC connection [30]. In this simulation, we omit the messages sent periodically from AGVs to the gNBs in order to provide information about the channel state. It can be seen that the 'RANDOM' baseline generates high signaling overhead, while the 'PPO-Game' and 'DDPG-Game' schemes result in the lowest signaling overhead. The signaling overhead of the 'PPO-Heuristic' scheme increases more rapidly than the 'PPO-Game' and 'DDPG-Game' schemes. This result confirms the CoMP clustering game is more efficient than the greedy heuristic user-centric CoMP clustering

VII. Conclusion

This paper has presented the joint CoMP clustering and beamforming problem for URLLC in an automated warehouse IIoT network. By combining a low complexity game-theoretic based CoMP clustering algorithm and the Proximal Policy Optimization method, we proposed an effective interference management framework that is suitable for a dynamic environment and can obtain performance approximated to the optimum and outperforms the user-centric CoMP clustering baseline.

References

[1] 3GPP, "Study on Communication for Automation in Vertical domains (Release 16)," in *TR 22.804 V16.2.0*, Dec. 2018. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3187

[2] P. Marsch and G. P. Fettweis, *Coordinated Multi-Point in Mobile Communications: from theory to practice*. Cambridge University Press, 2011.

[3] M. Khoshnevisan, V. Joseph, P. Gupta, F. Meshkati, R. Prakash, and P. Tinnakornsrisuphap, "5G industrial networks with CoMP for URLLC and time sensitive network architecture," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 947–959, 2019.

[4] A. A. Nasir, H. Tuan, H. H. Nguyen, M. Debbah, and H. V. Poor, "Resource Allocation and Beamforming Design in the Short Blocklength Regime for URLLC," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 2, pp. 1321–1335, 2020.

[5] P. Yang, X. Xi, Y. Fu, T. Q. Quek, X. Cao, and D. Wu, "Multicast eMBB and bursty URLLC service multiplexing in a CoMP-enabled RAN," *IEEE Transactions on Wireless Communications*, 2021.

[6] J. Khan and L. Jacob, "Resource Allocation for CoMP Enabled URLLC in 5G C-RAN Architecture," *IEEE Systems Journal*, 2020.

[7] P. Yang, X. Xi, T. Q. Quek, J. Chen, X. Cao, and D. Wu, "How should i orchestrate resources of my slices for bursty urllc service provision?" *IEEE Transactions on Communications*, 2020.

[8] J. Khan and L. Jacob, "Availability Maximization Framework for CoMP Enabled URLLC With Short Packets," *IEEE Networking Letters*, vol. 2, no. 1, pp. 1–4, 2020.

[9] P. Yang, X. Xi, T. Q. Quek, J. Chen, X. Cao, and D. Wu, "Ran slicing for massive iot and bursty urllc service multiplexing: Analysis and optimization," *IEEE Internet of Things Journal*, 2021.

[10] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1028–1042, 2020.

[11] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5g networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, 2019.

[12] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep Reinforcement Learning for Distributed Dynamic MISO Downlink-Beamforming Coordination," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6070–6085, 2020.

[13] L. Wang, G. Peters, Y.-C. Liang, and L. Hanzo, "Intelligent User-Centric Networks: Learning-Based Downlink CoMP Region Breathing," *IEEE Trans. Veh. Techno.*, vol. 69, no. 5, pp. 5583–5597, 2020.

[14] D. Shi, H. Gao, L. Wang, M. Pan, Z. Han, and H. V. Poor, "Mean field game guided deep reinforcement learning for task

placement in cooperative multiaccess edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9330–9340, 2020.
[15] D. Chen, Q. Qi, Z. Zhuang, J. Wang, J. Liao, and Z. Han, "Mean field deep reinforcement learning for fair and efficient uav control," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 813–828, 2020.
[16] Qualcomm, "5G: Bringing precise positioning to the connected intelligent edge," December 2021. [Online]. Available: https://www.qualcomm.com/media/documents/files/5g-positioning-for-the-connected-intelligent-edge.pdf
[17] M. K. Simon and M.-S. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," *Proceedings of the IEEE*, vol. 86, no. 9, pp. 1860–1877, 1998.
[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Info. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
[19] J. Wang, Z. Lan, C.-S. Sum, C.-W. Pyo, J. Gao, T. Baykas, A. Rahman, R. Funada, F. Kojima, I. Lakkis *et al.*, "Beamforming codebook design and performance evaluation for 60ghz wideband wpans," in *2009 IEEE 70th Vehicular Technology Conference Fall*. IEEE, 2009, pp. 1–6.
[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[21] F. Guidolin, L. Badia, and M. Zorzi, "A distributed clustering algorithm for coordinated multipoint in lte networks," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 517–520, 2014.
[22] M. M. Abdelhakam, M. M. Elmesalawy, K. R. Mahmoud, and I. I. Ibrahim, "A cooperation strategy based on bargaining game for fair user-centric clustering in cloud-ran," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1454–1457, 2018.
[23] A. F. Hanif, H. Tembine, M. Assaad, and D. Zeghlache, "Mean-field games for resource sharing in cloud-based networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 624–637, 2015.
[24] S. Ishikawa, "Fixed points by a new iteration method," *Proceedings of the American Mathematical Society*, vol. 44, no. 1, pp. 147–150, 1974.
[25] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (ddpg)-based energy harvesting wireless communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8577–8588, 2019.
[26] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for secure URLLC in mission-critical IoT scenarios," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5793–5807, 2020.
[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
[28] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[29] T. M. Ho, K.-K. Nguyen, and M. Cheriet, "UAV control for wireless service provisioning in critical demand areas: A deep reinforcement learning approach," *IEEE Trans. Veh. Techno.*, vol. 70, no. 7, pp. 7138–7152, 2021.
[30] 3GPP, "3GPP Radio Resource Control (RRC) protocol specification (Release 17) ," in *TS 38.331 V17.0.0 (2022-03)*, Mars 2022. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197

Tai Manh Ho received the B.Eng. and M.S. degrees from the Hanoi University of Science and Technology, Vietnam, in 2006 and 2008, respectively, and the Ph.D. degree from Kyung Hee University, South Korea, in 2018, all in computer engineering. Since 2019, he has been with the Synchromedia Lab, École de technologie supérieure, Université du Québec, Montréal, QC, Canada where he is currently a postdoctoral fellow. His current research interests include radio resource management and enabling technologies for 5G wireless systems.

Kim Khoa Nguyen is Associate Professor in the Department of Electrical Engineering at the University of Quebec's Ecole de technologie supérieure (ETS), Montreal, Canada. He has a Ph.D. from Concordia University in Electrical and Computer Engineering. In the past, he served as CTO of Inocybe Technologies (now is Kontron Canada), a world's leading company in software-defined networking (SDN) solutions. He was the architect of the Canarie's GreenStar Network and also involved in establishing CSA/IEEE standards for green ICT. He has led R&D in large-scale projects with Ericsson, Ciena, Telus, InterDigital, and Ultra Electronics. He is the recipient of Microsoft Azure Global IoT Contest Award 2017, and Ciena's Aspirational Prize 2018. His expertise includes cloud computing, network optimization, IoT, big data, machine learning, AI, smart building, smart city, high speed networks, and green ICT.

Dr. Mohamed Cheriet received his M.Sc. and Ph.D. degrees in Computer Science from the University of Pierre & Marie Curie (Paris VI) in 1985 and 1988 respectively. Since 1992, he has been a professor in the Systems Engineering department at the University of Quebec - École de Technologie Supérieure (ÉTS), Montreal, and was appointed full Professor there in 1998. Prof. Cheriet is the founder and director of Synchromedia Laboratory for multimedia communication in telepresence applications, since 1998.
Dr. Cheriet research has extensive experience in cloud computing and network virtualization and softwarisation. In addition, Dr. Cheriet is an expert in Computational Intelligence, Pattern Recognition, Machine Learning, Artificial Intelligence and Perception. Dr. Cheriet has published more than 450 technical papers in the field. He serves on the editorial boards of several renowned journals and international conferences. He is a holder in 2013 of a Tier 1 Canada Research Chair on Sustainable and Smart Eco-Cloud; he leads since then the establishment of the first smart university campus in Canada, created as a hub for innovation and productivity at Montreal. Dr. Cheriet is the General Director of the FRQNT Strategic Cluster on the operationalization of sustainability development, CIRODD (2019-2026). Dr. Cheriet is a 2016 Fellow of the International Association of Pattern Recognition (IAPR), a 2017 Fellow of the Canadian Academy of Engineering (CAE), a 2018 Fellow of the Engineering Institute of Canada (EIC), and a 2019 Fellow of Engineers Canada (EC). Dr. Cheriet is the recipient of the 2016 IEEE J.M. Ham Outstanding Engineering Educator Award, the ÉTS Research Excellence prize in 2013, for his outstanding contribution in green ICT, cloud computing, and big data analytics research areas, and of the 2012 Queen Elizabeth II Diamond Jubilee Medal. He is a senior member of the IEEE, the founder and former Chair of the IEEE Montreal Chapter of Computational Intelligent Systems (CIS), and a Steering Committee Member of the IEEE Sustainable ICT Initiative and the Chair of ICT Emissions Working Group. He published with his Working Group the first standard ever, IEEE 1922.2 on real-time calculation of ICT emissions, in April 2020.