Game Theoretic Reinforcement Learning Framework For Industrial Internet of Things

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

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771864

***Abstract*—The fifth-generation (5G) wireless network provides high-rate, ultra-low latency, and high-reliability connections that can meet the industrial IoT requirements in factory automation especially for swarm robotics communication. 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 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 planed reference tracks. Traditional iterative optimization approaches are impractical in such dynamic wireless environments 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, swarm robotics, industrial automation, URLLC, coordinated multipoint**

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 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 as an automated warehouse with high mobility swarm robotics, i.e., 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. In the automated warehouse scenario, CoMP can combine the signal 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.

To overcome the shortcomings of traditional optimization theory, recent works have proposed to use deep reinforcement learning (DRL) to address important aspects in CoMP communication such as clustering and beamforming design. In [3], a hybrid DRL model combining a deep deterministic policy gradient (DDPG) and a deep double Q-network (DQN) model is proposed to cluster the access points and optimize the beamforming vectors to maximize the sum rate. In [4] 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 [5], 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 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 [6], [7]. Shi *et al.* [6] 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 of the problem modeled by MFG in [7].

Unlike the existing works that consider 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. More specifically, we propose a distributed low complexity game-theoretic CoMP clustering algorithm combined with the Proximal Policy Optimization method to obtain a stationary solution for beamforming design for URLLC transmission between the gNBs and the AGVs in a highly dynamic environment of an automated warehouse application.

II. System Model and Problem Formulation

We consider an automated warehouse 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. 1). 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 with the minimum number of gNBs which can provide 5G communications with the required reliability to AGV k. Note that, these CoMP clusters can

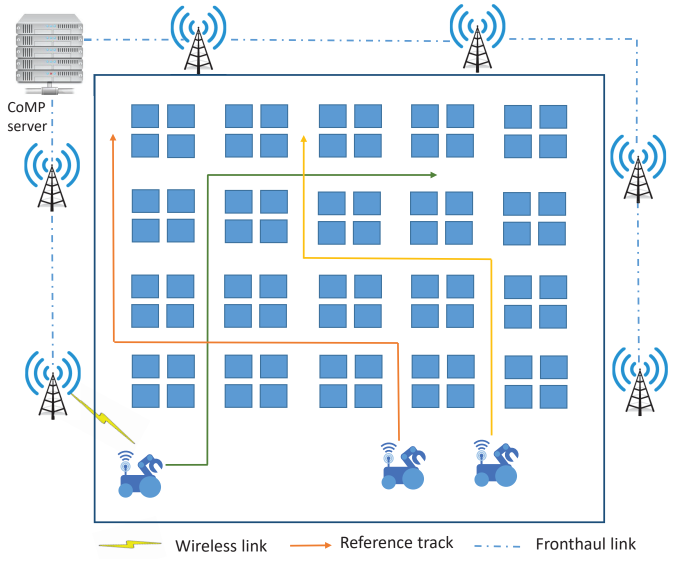

Figure 1: System model.

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 to simultaneously serve all AGVs within the warehouse area. We assume all the gNBs are deployed on the ceiling of the warehouse. Let $\mathbf{q}_{\mathrm{gNB},j} = [x_{\mathrm{gNB},j}, y_{\mathrm{gNB},j}]$ denote the coordinate of the gNB j and $z_{\mathrm{gNB},j}$ is the height of the gNB j. $\mathbf{q}_k[t] = [x_k[t], y_k[t]]$ represent the spatial coordinates of the AGV k. The distance between the gNB j and AGV k at time instant t is

$$d_{k,j}[t] = \sqrt{\left\|\mathbf{q}_{\mathrm{gNB},j} - \mathbf{q}_k[t]\right\|^2 + z^2_{\mathrm{gNB},j}} \tag{1}$$

A. Communication Model

The channel state information (CSI) is assumed to be estimated by the CoMP through training the pilot sequences. 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

$$\mathrm{y}_k = \underbrace{\sum_{j=1}^{B_k} \mathbf{h}^H_{k,j}\mathbf{w}_{k,j}\mathrm{s}_k}_{\text{Desired signal}} + \underbrace{\sum_{k'\neq k}^{K}\sum_{j=1}^{B_{k'}} \mathbf{h}^H_{k,j}\mathbf{w}_{k',j}\mathrm{s}_{k'}}_{\text{Interference}} + \sigma_k, \tag{2}$$

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

$$g_{k,j} = \left(\frac{c}{4\pi f_c}\right)^2 \left(\frac{d_{k,j}}{d_0}\right)^{-\alpha_g}, \tag{3}$$

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 [8]. 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 [8]

$$f(z, m) = \frac{2m^m}{\Gamma(m)\Omega^m} z^{2m-1} \exp\left(-\frac{m}{\Omega}z^2\right), \tag{4}$$

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]:

$$\gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \frac{\left|\sum_{j=1}^{B_k} \mathbf{h}^H_{k,j}\mathbf{w}_{k,j}\right|^2}{\sum_{k'\neq k}^{K}\left|\sum_{j=1}^{B_{k'}} \mathbf{h}^H_{k,j}\mathbf{w}_{k',j}\right|^2 + \sigma_k^2}, \tag{5}$$

$$\tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \log_2\Big(1 + \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j})\Big). \tag{6}$$

The maximum transmission rate to transmit D_k bits over n_k complex symbols in finite blocklength regime can be accurately approximated as [9]:

$$R_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) = \tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) - \sqrt{\frac{V}{n_k}}\frac{Q^{-1}(\epsilon_k)}{\ln(2)} \geq \frac{D_k}{n_k}, \tag{7}$$

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

$$\epsilon_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \leq Q\left(\frac{\ln(2)\left(\tilde{R}_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) - \frac{D_k}{n_k}\right)}{\sqrt{\frac{V}{n_k}}}\right). \tag{8}$$

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.

According to (8) and (5), the mathematical expression of the require beamformers $\{\mathbf{w}_{k,j}\}$ from the gNBs to AGV k that satisfies the decoding error probability ϵ_k requirements can be written as

$$\gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k), \tag{9}$$

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

$$\gamma_k^{th}(\epsilon_k) = \exp\left(\frac{D_k \ln(2)}{n_k} + \sqrt{\frac{V}{n_k}} Q^{-1}(\epsilon_k)\right) - 1. \quad (10)$$

B. Problem Formulation

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 difficult to design joint clustering and beamforming actions for all AGVs that consist of multiple matrices of continuous complex variables. In this paper, we propose using the codebook technique [4], [10], [11] so that the DRL 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 [4]:

$$\mathbf{w}_{k,j}[t] = \sqrt{p_{k,j}[t]}\bar{\mathbf{w}}_{k,j}[t], \quad (11)$$

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 (13c), 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 [4]

$$c_{m,q} = \frac{1}{\sqrt{M}} \exp\left(i\frac{2\pi}{\Phi} \left\lfloor \frac{m\text{mod}(q + \frac{Q_{\text{code}}}{2}, Q_{\text{code}})}{Q_{\text{code}}/\Phi} \right\rfloor \right), \quad (12)$$

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.

We consider the joint problem of CoMP clustering and beamforming design with the objective of maximizing the sum-rate across the AGVs subject to the URLLC constraint. To be specific, the joint problem in time slot t can be formulated as follows: **P1:**

$$\max_{\{\mathcal{B}_k\},\{p_{k,j}\},\{\bar{\mathbf{w}}_{k,j}\}} \sum_{k \in \mathcal{K}} R_k(\mathbf{w}_{k,j}[t], \mathbf{h}_{k,j}[t]) \quad (13a)$$

$$\text{subject to: } \gamma_k(\mathbf{w}_{k,j}, \mathbf{h}_{k,j}) \geq \gamma_k^{th}(\epsilon_k),\ \forall k \in \mathcal{K}, \quad (13b)$$

$$\sum_{k \in \mathcal{K}_j} p_{k,j}[t] \leq P_j, \forall j \in \mathcal{B}, \quad (13c)$$

$$\mathcal{B}_k[t] \subset \mathcal{B}, \forall k \in \mathcal{K}, \quad (13d)$$

$$\bar{\mathbf{w}}_{k,j}[t] \in \mathcal{C}, \forall k \in \mathcal{K}, \forall j \in \mathcal{B}_k[t], \quad (13e)$$

Constraint (13b) guarantees the reliability communication of AGV k, whereas (13c) sets a constraint on the total transmit power of gNB j. It can be seen that the problem **P1** in (13) is nonconvex combinatorial dues to the nonconvex objective function (13a) and the URLLC constraint (13b) and the combinatorial constraint (13d). This kind of problem is NP-hard. Therefore, we design a reinforcement learning framework to solve the problem by modeling the beamforming design problem as a Markov Decision Process (MDP).

III. Proximal Policy Optimization

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 **P1** 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 for each AGV at time t is designed as follow:

$$r_k[t] = \kappa_1 R_k[t] - \kappa_2 \sum_{j \in \mathcal{B}_k} P_{k,j}, \quad (14)$$

where the second term is the penalty for the gNBs in the cluster $\mathcal{B}_k$ and κ_1 and κ_2 are a turnable scale coefficients.

B. PPO-based Algorithm

Proximal policy optimization (PPO) [12] 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

$$\theta_{n+1} = \arg\max_{\theta} \underset{s,a \sim \pi_{\theta_n}}{\mathrm{E}} \left[L(s, a, \theta_k, \theta)\right], \quad (15)$$

where L is given in (16). $\pi_\theta(a|s)$ is new parameterized policy trying to seek the optimal parameter vector θ, and $\pi_{\theta_n}(a|s)$ is the old policy. 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

$$A^{\pi_{\theta_n}}(s, a) = Q^{\pi_{\theta_n}}(s, a) - V^{\pi_{\theta_n}}(s), \quad (19)$$

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.

$$Q^{\pi_{\theta_n}}(s_t, a_t) = \mathbb{E}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]. \quad (20)$$

The PPO algorithm is presented in Algorithm 1 and Fig. 2.

$$L(s,a,\theta_k,\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_n}(a|s)} A^{\pi_{\theta_n}}(s,a),\ \ \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_n}(a|s)}, 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_n}}(s,a)\right), \tag{16}$$

$$\theta_{n+1} = \arg\max_\theta \frac{1}{|\mathcal{D}_n|} \sum_{\tau\in\mathcal{D}_n} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t,a_t), g(\epsilon, A^{\pi_{\theta_k}}(s_t,a_t))\right); \tag{17}$$

$$\boldsymbol{\phi}^{(n+1)} = \arg\min_\phi \frac{1}{|\mathcal{D}_n|\Delta T} \sum_{\tau\in\mathcal{D}_n} \sum_{t=0}^{\Delta T} \left[V_{\boldsymbol{\phi}^{(n)}}(\boldsymbol{s}[t]) - r(\boldsymbol{s}[t], \boldsymbol{a}[t])\right]^2; \tag{18}$$

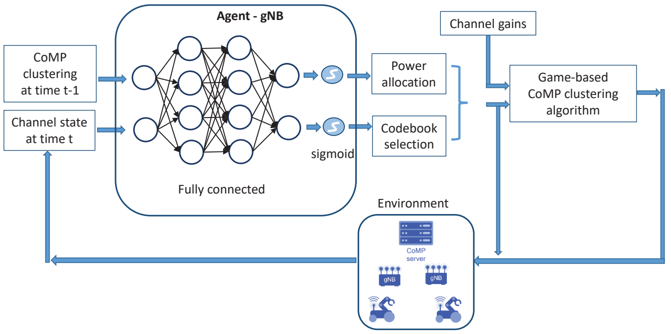

Figure 2: Joint CoMP clustering and beamforming design framework.

Algorithm 1 PPO-based beamforming design

```
Initialize policy and value function parameters $\theta^{(0)}$, $\phi^{(0)}$;
for $n = 0, 1, 2, ..$ iterations do
    The gNBs collect minibatches of $D$ transitions $\mathcal{D}_n = \{s_i, a_i, r_i, s_{i+1}\}_{i=0:D-1}$ by running policy $\pi_\theta$;
    Compute advantage estimates $\hat{A}(s_t, a_t)$ based on the current value function $V_{\phi^{(n)}}(s_t)$;
    Update the policy by maximizing the PPO-clip objective in (17) where
        $g(\epsilon, A) = \begin{cases} (1+\epsilon)A & A \geq 0 \\ (1-\epsilon)A & A < 0. \end{cases}$    (21)
    Fit value function by regression on MSE in (18)
end for
```

IV. Game Theoretic-Based CoMP Clustering

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 $a_{k,j} = 1$ if AGV k selects gNB j, and 0 otherwise. 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 $\boldsymbol{A} = \{\boldsymbol{a}_j\}_{j\in\mathcal{B}}$, $\boldsymbol{a}_j = \{a_{k,j}\}_{k\in\mathcal{K}, j\in\mathcal{B}}$ to maximize its payoff function.*
3) *Payoff function: The payoff of player k is given by*

$$\mathcal{P}_k(\boldsymbol{A}) = \sum_{j\in\mathcal{B}} \mathcal{P}_{k,j}(\boldsymbol{a}_j), \tag{22}$$

$$\begin{aligned}\mathcal{P}_{k,j}(\boldsymbol{a}_j) = &\frac{\left(a_{k,j}\left|\boldsymbol{w}_{k,j}\boldsymbol{h}_{k,j}^H\right|^2\right)^\alpha}{\sum_{l\in\mathcal{K}}\left(a_{l,j}\left|\boldsymbol{w}_{l,j}\boldsymbol{h}_{l,j}^H\right|^2\right)^\alpha} \\ &- \xi a_{k,j} \sum_{l\neq k\in\mathcal{K}} \left|\boldsymbol{w}_{k,j}\boldsymbol{h}_{l,j}^H\right|^2,\end{aligned} \tag{23}$$

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 [13].

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:

$$m_j = \left(\frac{1}{K}\sum_{k\in\mathcal{K}}\left(\omega_{k,j}a_{k,j}\right)^\alpha\right)^{\frac{1}{\alpha}}, \forall j \in \mathcal{B}. \tag{24}$$

The payoff function in (23) can be rewritten as follows:

$$\mathcal{P}_{k,j}(a_{k,j}, m_{j,-k}) = \frac{1}{K}\left(\frac{\omega_{k,j}a_{k,j}}{m_j}\right)^\alpha - \xi\mathcal{I}_{k,j}a_{k,j}, \tag{25}$$

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

The payoff function of a player has the following properties: i) The payoff function depends only on the player's action $a_{k,j}$ and the mean field m_j; and ii) 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*

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

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

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

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

$$\boldsymbol{Br}(a_{k,j}, m_j) = \underset{a_{k,j}}{\arg\max} \left[\frac{1}{K}\left(\frac{\omega_{k,j}a_{k,j}}{m_j}\right)^{\alpha} - \xi\mathcal{I}_{k,j}a_{k,j}\right], \tag{26}$$

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

$$a_{k,j}(\tau+1) = \lambda(\tau)\boldsymbol{Br}(a_{k,j}(\tau), m_j(\tau)) + (1-\lambda(\tau))a_{k,j}(\tau), \tag{27}$$

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

$$\boldsymbol{Br}(a_{k,j}(\tau), m_j(\tau)) = \left[K\left(m_j^{\alpha}(\tau) - \frac{K\xi\mathcal{I}_{k,j}m_j^{2\alpha}(\tau)}{\alpha\omega_{k,j}^{\alpha}a_{k,j}^{\alpha-1}(\tau)}\right)\right]^{\frac{1}{\alpha}}, \tag{28}$$

$$m_j(\tau+1) = \lambda(\tau)\left[\frac{a_{k,j}^{\alpha}(\tau)}{K} + \frac{K\xi\mathcal{I}_{k,j}m_j^{2\alpha}(\tau)}{\alpha\omega_{k,j}^{\alpha}a_{k,j}^{\alpha-1}(\tau)}\right]^{\frac{1}{\alpha}} + (1-\lambda(\tau))m_j(\tau). \tag{29}$$

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

The iterative best response update (27) has the form of Ishikawa (Mann) iteration [14]. It was proven in [14] 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 (27) converges strongly to a fixed point, which is a unique Nash equilibrium. □

A distributed game-based CoMP clustering is presented in Algorithm 2. 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.

Algorithm 2 Distributed Game-based CoMP Clustering

```
Initialize $a_{k,j}(0)$ and $m_j(0)$ $\forall k \in \mathcal{K}, j \in \mathcal{B}$;
All gNBs broadcast their beamforming profiles $\mathbf{w}_{k,j}$;
repeat
    Each AGV $k$ updates its strategy $a_{k,j}(\tau)$ according to (27) and (28);
    Update mean field according to (29);
until $|a_{k,j}(\tau+1) - a_{k,j}(\tau)| \leq \varepsilon$
```

V. Simulation Results

We perform extensive simulations to evaluate the performance of our proposed design in terms of the sum URLLC rate in (7), i.e., the objective of the optimization problem **P1**. We vary the number of AGVs from 2 - 12, 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. 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 standard deviation of 2. The carrier frequency 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.

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 [15] combined with the distributed game theoretic based CoMP clustering Algorithm 2. We investigate whether on-policy or off-policy outperforms in a dynamic environment as in a robotic network.
- 'PPO-GREEDY': A greedy CoMP clustering algorithm where each AGV selfishly searches a set of serving gNBs based on its received signal strength without considering interference caused to other AGVs.
- 'EXHAUST': We use the exhaustive search method over the Euclidean space $\mathcal{B}\times\mathcal{C}\times P$. The 'EXHAUST' baseline is considered the optimal solution for the formulated problem.

Fig. 3 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 'PPO-GREEDY' 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-GREEDY' baseline and reaches a stable reward close to the 'EXHAUST' baseline which is the optimum. Moreover, the 'DDPG-Game' baseline has a similar conver-

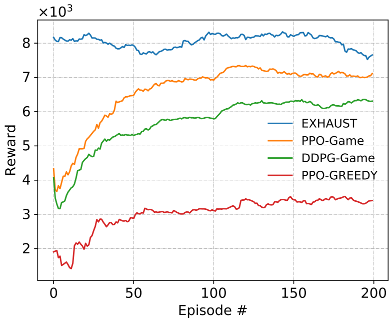

Figure 3: Accumulative reward

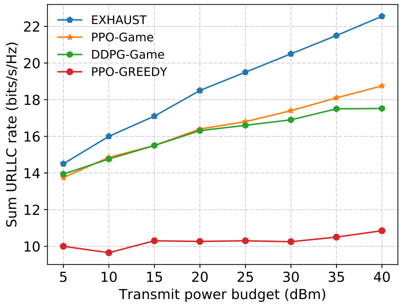

Figure 4: Sum URLLC rate

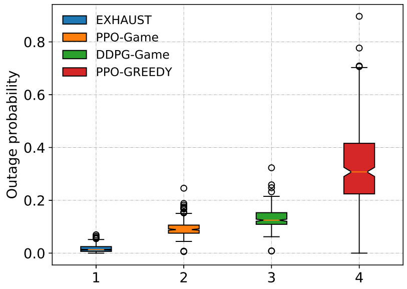

Figure 5: Outage probability

gence behavior but converges to a lower value compared to the ‘PPO-Game’ scheme.

Fig. 4 plots the sum URLLC rates versus the transmit power budget. It can be seen that the sum URLLC rate 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-GREEDY’ baseline is nearly constant. This result again confirms the ‘PPO-Game’ scheme can manage interference better than the ‘PPO-GREEDY’ baseline. When the transmit power increases the interference also increases, hence, an interference adaptive scheme would be beneficial.

Fig 5 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%. The outage probability of the ‘PPO-Game’ scheme is lower than that of the ‘DDPG-Game’ scheme, at around 9% and 13%, respectively. The ‘PPO-GREEDY’ 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%.

VI. Conclusion

This paper has presented the joint CoMP clustering and beamforming problem for URLLC in an automated warehouse 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 greedy heuristic CoMP clustering baseline.

Acknowledgment

The authors thank Mitacs, Ciena, and ENCQOR for funding this research under the grant IT13947

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] 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 Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1028–1042, 2020.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[13] 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.

[14] S. Ishikawa, “Fixed points by a new iteration method,” *Proceedings of the American Mathematical Society*, vol. 44, no. 1, pp. 147–150, 1974.

[15] 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.