Age-of-Information Aware Scheduling for Edge-Assisted Industrial Wireless Networks

Mingyan Li, *Member, IEEE*, Cailian Chen, *Member, IEEE*, Huaqing Wu, *Student Member, IEEE*, Xinping Guan, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Industrial wireless networks (IWNs) have attracted significant attention for providing time-critical delivery services, which can benefit from device-to-device (D2D) communication for low transmission delay. In this article, a distributed scheduling problem is investigated for D2D-enabled IWNs, where D2D links have various age-of-information (AoI) constraints for information freshness. This problem is formulated as a constrained optimization problem to optimize D2D packet delivery over limited spectrum resources, which is intractable since D2D users have no prior knowledge of the operating environment. To tackle this problem, in this article, an AoI-aware scheduling scheme is proposed based on primal-dual optimization and actor–critic reinforcement learning. In specific, multiple local actors for D2D devices learn AoI-aware scheduling policies to make on-site decisions with their stochastic AoI constraints addressed in the dual domain. An edge-based critic estimates the performance of all actors' decision-making policies from a global view, which can effectively address the nonstationary environment caused by concurrent learning of multiple local actors. Theoretical analysis on the convergence of learning is provided and simulation results demonstrate the effectiveness of the proposed scheme.**

***Index Terms*—Age of information (AoI), device-to-device (D2D) communication, edge computing, industrial wireless network (IWN), reinforcement learning.**

I. Introduction

INDUSTRIAL wireless networks (IWNs) [1] have gained significant attention in smart factory due to the advantages of flexibility, low cost, and scalability. For delay-sensitive industrial applications, e.g., state perception and estimation, IWNs are also envisioned to guarantee the service timeliness, which is crucial since missing a deadline may lead to production inefficiency, equipment destruction, and so on [2]. As a typical industrial application, monitoring systems have stringent requirements for information freshness, which can be quantified by the newly proposed age-of-information (AoI) metric (also referred to as age) [3]. As a novel metric of timeliness, AoI measures the time elapsed since the generation of the most recently delivered packet from the perspective of destinations and is paramount in a wide range of information, communication, and control systems. Therefore, AoI should be stringently bounded in IWNs.

As a promising cellular-based technique for proximity transmissions, device-to-device (D2D) communication specially fits to the industrial monitoring systems where numerous nearby sensor–actuator pairs require timely state update [4]. Each D2D device pair can communicate directly with little involvement of the base station (BS), resulting in extremely low power consumption and delays. However, due to the limited wireless resources named resource blocks (RBs) in cellular networks, the available RBs should be intelligently allocated especially for large-scale IWNs to coordinate the competition of numerous D2D links [5], [6]. Therefore, how to schedule D2D packet delivery over limited RBs while guaranteeing AoI constraints is a critical problem in IWNs.

Recently, age-aware link scheduling has received significant attention since AoI was first proposed in [7]. Specifically, the key idea in [8]–[13] is to reduce the average AoI of users via a single shared wireless channel. Focusing on IWNs, it is more profitable to address age constraints rather than average age due to the determination requirements of industrial applications [14], [15]. Considering age constraints, an energy efficiency problem is studied in [16] with global channel state information (CSI). In D2D networks, however, the acquisition of global CSI is challenging [17].[1] In addition, most existing works on age are designed in a centralized manner, which may incur large latency and scalability issues. Although distributed scheduling can enhance network scalability and is more suitable for D2D-enabled IWNs, the scheduling strategies of D2D device pairs will be

Manuscript received June 19, 2020; revised September 23, 2020; accepted October 29, 2020. Date of publication November 11, 2020; date of current version May 3, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1702100, in part by the Natural Science Foundation of China under Grant 61933009 and Grant 61622307, in part by the Program of Shanghai Academic Research Leader under Grant 19XD1421800, and in part by the China Scholarship Council under Grant 201906230168. Paper no. TII-20-2988. *(Corresponding author: Cailian Chen.)*

Mingyan Li, Cailian Chen, and Xinping Guan are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China (e-mail: limy2015@sjtu.edu.cn; cailianchen@sjtu.edu.cn; xpguan@sjtu.edu.cn).

Huaqing Wu and Xuemin Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: h272wu@uwaterloo.ca; sshen@uwaterloo.ca).

Color versions of one or more of the figures in this article are available at https://doi.org/10.1109/TII.2020.3037299.

Digital Object Identifier 10.1109/TII.2020.3037299

[1] The inband mode of D2D communication [18] is adopted in this article, where D2D pairs can communicate directly via their LTE-A interface. As a result, conventional pilot signals sent by the BS cannot be utilized for channel estimation as D2D packets will not be forwarded across the BS.

myopic in a fully distributed environment since they have only local observations.

Different from the previous works, we consider the age-aware link scheduling problem under the interference-limited access model.[2] In such a problem, D2D devices aim to share multiple RBs in order to satisfy their age constraints without requiring prior knowledge of channel quality. Moreover, an edge-assisted hierarchical network is considered, which is shown to have significant advantages in efficiency of resource management and is widely adopted in IWNs [20]. In such a network, edge controllers can manage the communication in IWNs to boost mission-critical and personalized edge intelligence in the future [21], [22]. This motivates us to design a novel scheme for age-aware distributed link scheduling, where edge controllers can assist distributed nodes to autonomously access the network so as to improve communication efficiency and reduce network pressure.

Leveraging edge computing, an age-aware scheduling scheme, namely AoIS, is proposed in this article for large-scale D2D-enabled IWNs, where multiple D2D links share spectrum in a distributed manner. The key idea of AoIS is to guide each D2D user to intelligently select an RB and transmit power level to deliver its state packets, so that the AoI constraint can be respected. Finding solutions to such a problem is a significant challenge. The reasons are threefold and given as follows:

1) this problem is a functional optimization problem with stochastic age constraints, which is intractable to solve directly;
2) D2D users are scheduled with local observation in a distributed manner, and thus do not have prior knowledge of the operating environment, i.e., the information on both channel quality and scheduling actions of others;
3) the environment is nonstationary due to the concurrently updated scheduling policies of D2D users.

To deal with these challenges the following statements hold.

1) The proposed AoIS first uses near-universal learning parameterizations to represent the scheduling strategy for each D2D user. Then, parameter training is undertaken in both primal and dual domains to transform the optimization formulation into an unconstrained learning problem.
2) Model-free reinforcement learning (RL) is used in AoIS so that parameters of each user can be updated without requiring prior knowledge of the operating environment.
3) To improve learning stability, AoIS utilizes the actor-critic framework where an edge controller serves as a critic to estimate the policies of D2D users from the global viewpoints so that local actors can make scheduling decisions with others' information taken into account.

The remainder of this article is organized as follows. Section II proposes the network architecture. Section III presents the system model and problem formulation. In Section IV, the primal-dual RL is formulated. Section V explains the details of AoIS. Section VI presents the performance analysis of AoIS. Finally, Section VII concludes this article.

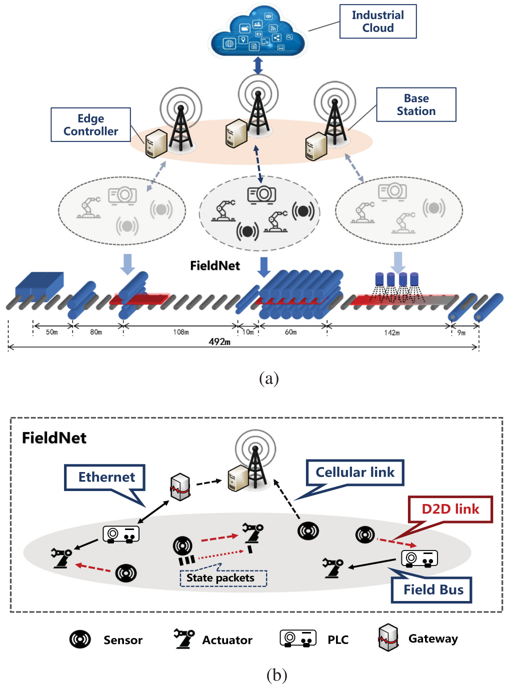


Fig. 1. Illustration of the hierarchical IWNs. (a) Edge-cloud hierarchical IWNs architecture. (b) Edge-assisted and D2D-enabled FieldNet.

II. Architecture of Hierarchical IWNs

A three-layer hierarchical industrial architecture is proposed in this work, shown in Fig. 1(a), which consists of field, edge, and cloud layers. The first layer is the field network layer, which contains substantial geography-distributed devices, such as programmable logic controllers, sensors, actuators, robots, etc. For flexible and scalable resource management, field networks can be divided into multiple small-scale networks, called FieldNets, according to different industrial subprocesses. For example, the hot rolling process (with the length of 492 m), as a typical example of large-scale industrial cyber–physical systems (CPSs), is composed of several subprocesses, such as reversing rougher, finishing milling, and laminar cooling as shown in Fig. 1(a). In each FieldNet, there exist many sensor nodes that monitor the state of field devices, industrial processes, and product quality, as well as an edge-based small BS managing the communication in this FieldNet. Hence, the second layer is the access layer including multiple small BSs and edge computing nodes with storage, computing, communication, and other resources. Since this layer is closer to field networks, it can be used for processing of computing-intensive tasks, e.g., analyzing industrial big data from field devices [23]. Moreover, an edge-based BS can deliver some data to the last layer, which is the industrial cloud, where historical data of field devices and edge nodes can be stored in the cloud for long-term data analysis.

In each FieldNet, conventional cellular links along with industrial Ethernet and field bus networks form a heterogeneous industrial network, as shown in Fig. 1(b). Moreover, D2D device pairs communicate directly in a distributed manner for the industrial wireless applications in need of timely state update, where AoI is a critical metric for information freshness [20]. In this article, we focus on age-aware D2D communication in

[2] In the interference-limited case, users can simultaneously transmit using the same RB with acceptable interference to each other [17], [19].

a FieldNet, where state packets are transmitted directly among D2D pairs with an edge-based BS assisting the communication. The resource management of different FieldNets with multiple edge controllers will be our future work.

III. System Model and Problem Statement

A. System Model

As shown in Fig. 1(b), we consider a distributed D2D-based sensing system in IWNs, composed of a set $\mathcal{N} = \{1, \ldots, N\}$ of transmitter–receiver pairs for state update under the coverage of a BS. Each device pair is referred to as a D2D user, which is denoted either by just n or by the pair (n, n'). All D2D pairs share a set $\mathcal{K} = \{1, \ldots, K\}$ of RBs with the bandwidth ω per RB. For link scheduling, a time-slotted system is considered, where scheduling decisions are made at the beginning of each time slot t. Each time slot has a duration of τ, which is equal to 1 ms or even smaller by using transmission time interval shortening technology. The beginning and ending instants of slots are synchronized across all nodes in the network.

In industrial CPSs, state update packets usually contain a small amount of information but require fresh data. For a D2D user n, the number of packets generated at time slot t can be denoted by $A_n(t)$, which follows a Poisson arrival process, specifically, M/G/1, with the average packet arrival rate per slot defined by κ_n. In M/G/1 systems, arrivals are Markovian (Poisson), which captures the scenario where the packet generation in IWNs is triggered by some random events [24].[3] State updates are generated and stored at transmitters' queues, following the first-come-first-serve (FCFS) policy. Denoting the D2D pair ns queue length at the beginning of time slot t by $Q_n(t)$, the queue dynamics without packet loss is given by

$$Q_n(t+1) = \max\{Q_n(t) - R_n(t), 0\} + A_n(t) \tag{1}$$

where $R_n(t)$ is the data rate of D2D user n at time slot t (in the unit of packets per slot), which will be derived in the next.

Let every packet be time stamped with the time it was generated. The AoI of D2D user n is defined as the time elapsed since the generation of the latest packet that has departed the transmitter given by

$$q_n(t) \triangleq t - \max_i \{t_A^n(i) \mid t_D^n(i) \leq t\} \tag{2}$$

where $t_A^n(i)$ and $t_D^n(i)$ are the arrival and departure instants of the ith packet of D2D user n. For timely delivery, each D2D user aims to satisfy its AoI constraint as

$$q_n(t) \leq d_n \quad \forall n \in \mathcal{N} \tag{3}$$

where d_n is the largest AoI that D2D pair n can support and the diversity of d_n captures the customized quality of service (QoS) level of each D2D user. Given the data rate $R_n(t)$, the evolution of AoI for D2D user n can be written as

$$q_n(t) = \begin{cases} q_n(t-1)+1 & , \text{ if } R_n(t) = 0 \\ t - \max_i\{t_A^n(i) \mid t_D^n(i) \leq t\} & , \text{ if } R_n(t) \neq 0 \end{cases} . \tag{4}$$

B. Communication Model

At the beginning of each time slot, every D2D user selects an RB and power level to transmit packets. For the sake of practical circuit restriction, the transmit power is discretized, where L power levels are considered. Given the power level $p_n(t) \in \{1, \ldots, L\}$, the instantaneous transmit power of each D2D user can be calculated as

$$P_n(t) = \frac{P_{\max}}{L} \cdot p_n(t) \quad \forall n \in \mathcal{N} \tag{5}$$

where $P_{\max}$ is the maximal allowable power. Since D2D users are allowed to access all the RBs, collision might occur. To serve massive devices in IWNs, we adopt the interference-limited access model [17], where interfered users can transmit together. Accordingly, the transmission data rate of D2D user n on RB k at time slot t is given by

$$R_n(t) = \frac{\omega\tau}{Z} \log_2 \left(1 + \frac{P_n(t) h_{nn'}^k(t)}{N_0\omega + \sum_{m \in \{\mathbf{C}_k \backslash n\}} P_m(t) h_{mn'}^k(t)}\right) \tag{6}$$

where Z is the packet length in bits. Moreover, $\mathbf{C}_k \subset \mathcal{N}$ is the set composed of D2D users operating over the same RB, called a D2D user coalition, and $h_{nn'}(t)$ is the instantaneous channel gain between devices n and n'.

C. Age-Aware Autonomous Scheduling Formulation

We aim to design an age-aware autonomous scheduling policy for each D2D user that optimizes its data rate performance, subject to its AoI constraint. Consider a scheduling policy π_n that returns a schedule $\phi_n := \{\boldsymbol{c}_n \in \{0,1\}^K, \boldsymbol{p}_n \in \{0,1\}^L\}$ defined by the channel selection vector $\boldsymbol{c}_n$ and power level selection vector $\boldsymbol{p}_n$ with $\sum_{k=1}^K c_{n,k} = 1$, $\sum_{j=1}^L p_{n,j} = 1$. The expected age-aware optimal scheduling formulation for D2D users is

$$\begin{aligned} J^* := \ & \max_{\pi_n} \ \mathbb{E}_{\mathbf{H},\mathbf{\Pi}}\left[R_n(t)\right] \\ & \text{s.t. } \mathbb{E}_{\mathbf{H},\mathbf{\Pi}}\left[q_n(t)\right] \leq d_n, \\ & \pi_n = \left\{\boldsymbol{c}_n \in \{0,1\}^K, \boldsymbol{p}_n \in \{0,1\}^L\right\} \end{aligned}$$

where $\mathbf{H}$ denotes the environment consisting of both global CSI (i.e., $h_{nm}(t)$ between any two D2D nodes) and packet arrivals of all D2D users (i.e., κ_n). $\mathbf{\Pi} = \{\pi\}^{-n}$ is the scheduling strategy matrix of others, determining the level of interference.

IV. Age-Aware Scheduling Based on Primal-Dual RL

The aim is to find suitable age-aware scheduling policies for a set of D2D users so that their data rate can be maximized with AoI constraints satisfied. However, this problem is generally difficult to solve [25] due to the following challenges.

1) The optimization problem contains stochastic constraints with the functional variable π_n.
2) The D2D user n has no prior information on not only the complete $\mathbf{H}$ but also other users' scheduling strategies $\mathbf{\Pi}$.

The first challenge indicates that J^* is a functional optimization problem with stochastic constraints, which is intractable to solve due to high computational complexity. Furthermore, the

[3] Event-triggered systems are considered in this article. Moreover, AoIS can also be applicable to time-triggered systems, where the given sampling periods can be served as a kind of input.

data rate in (6) and AoI evolution in (4) are unknown before D2D users take their scheduling actions due to the second challenge.

A. Deep Learning Formulation

To deal with the functional variable, we use the statistical learning, which is accomplished by introducing a parameterization of the scheduling function $\pi_n(o_n, \boldsymbol{\theta}_n)$, defined with the parameter $\boldsymbol{\theta}_n \in \mathbb{R}^x$ of some finite-dimensional x, where o_n is the local observation of D2D user n. In the proposed age-aware autonomous scheduling problem, D2D users make decisions according to only local observation o_n and their priority is to satisfy their AoI constraints. Therefore, o_n should reflect the possibility of AoI violation. To this end, the following lemma regarding the AoI constraint is proposed.

Lemma 1: For an M/G/1 queuing system following FCFS policy with an age deadline d_n, the AoI constraint of D2D user n in (3) will be satisfied when

$$Q_n(t) < R_n(t) + \tilde{A}_n(t - d_n) \tag{7}$$

where $\tilde{A}_n(t - d_n) + 1$ is the number of state update packets arriving during the time slots $[t - d_n, t)$, which can be observed in the D2D transmitter n.

Proof : At time slot t, let i_e be the index of the packet at the end of queue and $\hat{i}$ be the index of the first arriving packet at or after time $t - d_n$. Then, the index of the scheduled packets in this slot can be written as [16]

$$i_e + 1 - Q_n(t) \leq i \leq i_e - \max\left(Q_n(t) - R_n(t), 0\right). \tag{8}$$

Then, let $t_D^n(\hat{i}) > t$ represent the event that $\hat{i}$ is not served before or at time t, which also denotes the violation of AoI constraint. Accordingly, we have

$$\begin{aligned} q_n(t) > d_n &\overset{(a)}{\leftrightarrow} t_D^n(\hat{i}) > t \\ &\overset{(b)}{\leftrightarrow} \hat{i} > i_e - \max\left(Q_n(t) - R_n(t), 0\right) \\ &\overset{(c)}{\leftrightarrow} Q_n(t) > R_n(t) + i_e - \hat{i} \end{aligned}$$

where (a) is based on the definition of AoI, (b) is based on (8), and (c) is obtained according to $t_D^n(\hat{i}) > t$. ■

Based on (7), the age debt of D2D user n is defined, which denotes the penalty for the violation of AoI constraints. That is

$$D_n(t) = Q_n(t) - \left(R_n(t) + \tilde{A}_n(t - d_n)\right). \tag{9}$$

This denotes the number of additional updates that a D2D user should deliver so that its age constraint can be met. By taking the age debt into account, the local observation can be defined as $o_n = \{D_n\}$. As a result, the optimization problem J^* can be rewritten as

$$\begin{aligned} J_\theta^* := \max_{\boldsymbol{\theta}_n} \ & \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}\left[f_R(\pi_n(o_n, \boldsymbol{\theta}_n))\right] \\ \text{s.t. } & \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}[Q_n(t) - (f_R(\pi_n(o_n, \boldsymbol{\theta}_n)) + \tilde{A}_n(t - d_n))] \leq 0 \\ & \pi_n(o_n, \boldsymbol{\theta}_n) = \{\boldsymbol{c}_n \in \{0,1\}^K, \boldsymbol{p}_n \in \{0,1\}^L\},\ \boldsymbol{\theta}_n \in \mathbb{R}^x \end{aligned}$$

where a scheduling action ϕ_n is made via the policy $\pi_n(o_n, \boldsymbol{\theta}_n)$, parameterized by $\boldsymbol{\theta}_n$, with the input of local observation o_n. Then, deep neural networks (DNNs) are employed to approximate the scheduling function. As a result, the optimization problem J_θ^* is performed over $\boldsymbol{\theta}_n$ rather than the scheduling policy directly. Moreover, we rewrite the data rate in the form of function as $f_R(\pi_n(o_n, \boldsymbol{\theta}_n))$, which partially depends on the scheduling output $\pi_n(o_n, \boldsymbol{\theta}_n) = \phi_n$.

B. Unconstrained Primal-Dual Learning

To solve J_θ^*, the weights of DNN need to be found in order to maximize the received data rate while satisfying the AoI constraint. However, the standard approach of gradient-based optimization methods of DNN cannot be applied directly due to the AoI constraint. Thus, an unconstrained formulation is needed to capture the form of J_θ^*. A naive penalty-based reformulation will introduce a similar but fundamentally different problem, so we opt for constructing a Lagrangian function for J_θ^*, given by

$$\begin{aligned} \mathcal{L}_\theta(o_n, \lambda_n, \boldsymbol{\theta}_n) &:= \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}[f_R(\pi_n(o_n, \boldsymbol{\theta}_n)) \\ &- \lambda_n(Q_n(t) - f_R(\pi_n(o_n, \boldsymbol{\theta}_n)) - \tilde{A}_n(t - d_n))]. \end{aligned} \tag{10}$$

This Lagrangian penalizes the users with AoI violation through the second term, i.e., the age debt $f_D(\pi_n(o_n, \boldsymbol{\theta}_n)) := Q_n(t) - f_R(\pi_n(o_n, \boldsymbol{\theta}_n)) - \tilde{A}_n(t - d_n)$, scaled by the dual parameter λ_n.

To deal with this Lagrangian dual problem, the saddle point formulation is defined as [25]

$$D_\theta^* := \min_{\lambda_n \geq 0} \max_{\boldsymbol{\theta}_n} \mathcal{L}_\theta(o_n, \lambda_n, \boldsymbol{\theta}_n). \tag{11}$$

This renders the primal J_θ^* analogous to conventional learning problems, which can be solved with gradient-based optimization algorithms. Since this parameterized formulation is nonconvex, there exists a duality gap, defined as the difference $J^* - D_\theta^*$ between the dual and primal optima. In fact, the duality gap can be very small when using the parameterization that is near-universal[4] according to [25, Th. 1]. Specifically, it shows that the duality gap is bounded by $\|\boldsymbol{\lambda}^*\|_1 L\epsilon$. $\|\boldsymbol{\lambda}^*\|_1$ is the multiplier norm, which is related to the assumption stating that service demands are provisioned with some slack. The constant L is introduced to guarantee Lipschitz continuity of scheduling policies. Although both assumptions restrict the scope of targeted problems, they still allow consideration of most communication problems of practical importance [25]. Besides, the factor ϵ comes from the error of approximating resource allocations using the parameterized policy $\pi_n(o_n, \boldsymbol{\theta}_n)$, and can be very small when using near-universal parameterizations, e.g., DNN. This result justifies the operation in the dual domain with DNN parameterization as it does not entail a significant loss of optimality.

Therefore, we use primal-dual learning to solve (11) in the dual domain. A primal-dual method performs gradient updates on both primal and dual variables of the Lagrangian function in (10) to find a local stationary point of the optimization

[4] The policy with near-universal parameterization can model any function π_n within a stated accuracy [26].

problem J_{θ}^{*}. Formally, we update the primal parameters $\boldsymbol{\theta}_n$ of scheduling policy and the dual parameter λ_n by successively moving them toward the maximum and minimum points of the Lagrangian function as in (11). At each iteration t, by adding the corresponding partial gradients of the Lagrangian $\nabla_{\boldsymbol{\theta}}\mathcal{L}$, we have

$$\lambda_{n,t+1} = [\lambda_{n,t} + \beta\mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}[f_D \pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t}))]]_+ \tag{12}$$

$$\begin{aligned}\boldsymbol{\theta}_{n,t+1} = \boldsymbol{\theta}_{n,t} &+ \alpha\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}[f_R(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t}))\\ &- \lambda_{n,t+1} f_D(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t})))]\end{aligned} \tag{13}$$

where $\beta > 0$ and $\alpha > 0$ are the step sizes of (12)-(13).

C. Primal-Dual Learning Based on Policy Gradient

Recall that D2D users do not have prior information on the operating environment, that is $\boldsymbol{\Pi}$ and $\mathbf{H}$. Therefore, D2D users do not have the analytic form of $f_R(\pi_n(o_n, \boldsymbol{\theta}_n))$ and the expected gradient in (13) cannot be computed. In this regard, model-free RL methods are adopted to calculate the gradient, where a stochastic policy, rewritten as $\pi_n\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle$, is constructed to decide scheduling decisions $\phi_{n,t}$ for D2D user n at each time slot t. Accordingly, the expected gradient can be inferred based on policy gradient [27], given by

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}[g(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t})]\\ &= \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}\left[g_{n,t}\nabla_{\boldsymbol{\theta}} \log \pi_n\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle\right]\\ &\quad g(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t})\\ &= f_R(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t})) - \lambda_{n,t+1} f_D(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t}))\\ &= (1 + \lambda_{n,t+1}) f_R(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t}))\\ &\quad - \lambda_{n,t+1}(Q_n(t) - \tilde{A}_n(t - d_n))\end{aligned} \tag{14}$$

which outputs the reward value $g_{n,t}$. In this way, the gradients can be obtained by interacting with the environment rather than analytic computation done via model knowledge.

Scheduling policies can be learned effectively by using the policy-based methods, which have good convergence properties. However, traditional policy-based methods sometimes are inefficient to evaluate a policy due to the large variance of reward estimation. Thus, value-based methods, such as Q-learning, are used to ameliorate high variance, however, lead to high bias. In this regard, actor-critic RL methods are proposed [28] to combine the process of policy-based and value-based algorithms, which are used in our distributed age-aware scheduling. The role of an actor is to define a parameterized policy and generate actions, whereas the critic is in charge of evaluating and criticizing the current policy by processing the rewards received from the environment.

V. AoIS: Age-Aware Scheduling Scheme for Distributed D2D-Enabled IWNs

A. Edge-Assisted Actor-Critic Learning in AoIS

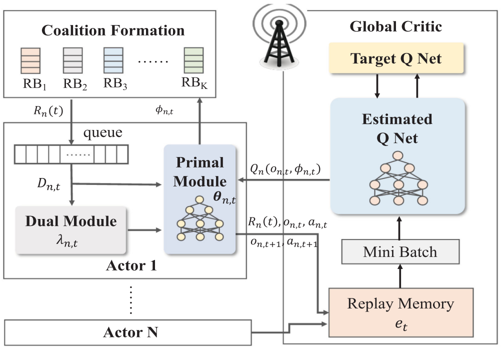


Fig. 2. Edge-assisted actor-critic learning framework.

In multiagent learning systems, state transitions of agents' common environment depend on their joint action. Thus, traditional actor-critic methods are still unsuitable in the setting with multiple RL agents due to the nonstationarity issue. To be specific, the rewards conditioned only on the user's own actions (when the actions of other users are not considered in the optimization process) exhibit much more variability, resulting in nonstationary environment. To stabilize the learning process, we propose an edge-assisted actor-critic learning framework by taking advantage of edge computing in IWNs, shown in Fig. 2. In this framework, AoIS is composed of two parts: local actors for N D2D users and an edge-based critic. The latter learns an approximation of action-state value function for local actors from the perspective of global coalition structure and its evolution. Based on the feedback of critic, the local actor updates its policy parameters and makes scheduling actions on-site. Therefore, the proposed learning framework features centralized estimation and decentralized execution.

Formally, the critic aims to estimate the action-state value function related to the data rate for each D2D user based on the current coalition structure. According to the definition of $g(\pi_n(o_{n,t}, \boldsymbol{\theta}_{n,t}))$ in (14), the action-state value function outputs Q values of expected long-term reward of $(1 + \lambda_{n,t+1})R_n(t)$, if the scheduling action $\phi_{n,t}$ is taken at time slot t_0 with the extended local observation $o_{n,t} = \{D_{n,t}, \lambda_{n,t}\}$ defined as

$$\begin{aligned}&\boldsymbol{Q}_n(o_{n,t}, \phi_{n,t})\\ &\quad = \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}\left[\sum_{t=t_o}^{T} \gamma^{t-t_o}(1 + \lambda_{n,t+1})R_n(t) \mid \mathbf{o}_t^{-n}, \boldsymbol{\phi}_t^{-n}\right]\end{aligned} \tag{15}$$

where γ is the discount factor and $\mathbf{o}_t^{-n}$ and $\boldsymbol{\phi}_t^{-n}$ denote the local observation and scheduling action vector of other users, respectively. Specifically, the edge-based critic aims to output Q values for each actor after scheduling, with the global coalition structure $s_t = \{o_{1,t}, \ldots, o_{N,t}\}$ and $a_t = \{\phi_{1,t}, \ldots, \phi_{N,t}\}$. $\boldsymbol{Q}_n(o_{n,t}, \phi_{n,t})$ in (15) can be rewritten as $\boldsymbol{Q}_n(s_t, a_t)$, which is trained centrally in the edge by collecting local observations, actions, and the corresponding data rates from each D2D user. Therefore, the critic is informed of the experience $e_t = (s_t, a_t, r_t, s_{t+1}, a_{t+1})$ where $r_t = (R_1(t), \ldots, R_N(t))$. s_{t+1}, a_{t+1} represent the coalition structure formed by the next scheduling. Then, the critic is trained with deep Q networks (DQN) [29], where the loss function $S(t)$ of $\boldsymbol{Q}_n(s_t, a_t)$ is proportional to temporal difference errors

$$S(t) = \mathbb{E}_{\mathbf{H},\boldsymbol{\Pi}}\left[\sum_{n\in\mathcal{N}} |\boldsymbol{Q}_n(s_t, a_t) - y_{n,t}|\right] \tag{16}$$

$$y_{n,t} = (1 + \lambda_{n,t+1})R_n(t) + \gamma\boldsymbol{Q}_n'(s_{t+1}, a_{t+1}) \text{ for } n \in \mathcal{N}. \tag{17}$$

Algorithm 1: The Workflow of AoIS.

```
Input: Scalar step size $\alpha$ and $\beta$; Learning rate of
       critic $\eta$, Batch size $\mathcal{B}$; Experience size $\mathcal{E}$;
       Initial $\lambda_{n,0}$ and $\boldsymbol{\theta}_{n,0}$ for each local actor $n$;
\* Workflow of Local Actor *\
for $t = 1, 2, 3 \ldots$ do
    1. Update age debt by adding up packet arrivals
       and get the local observation $o_{n,t} = \{D_{n,t}, \lambda_{n,t}\}$ ;
    2. Take the scheduling action based on the output
       of policy $\pi_n\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle$ ;
    3. Get data rate $R_n(t)$ and update age debt $D_{n,t}$ by
       Eqn. (9) and update the dual variable based on
           $\lambda_{n,t+1} = \mathrm{P}_\Lambda\left[\lambda_{n,t} + \beta D_{n,t}\right]$;
    4. Deliver $R_n(t)$ along with $o_{n,t}, o_{n,t+1}$ and $\phi_{n,t}, \phi_{n,t+1}$
       to the critic for experience generation;
    5. Receive $Q_n(o_{n,t}, \phi_{n,t})$ from critic and then
       generate a sample $\{o_{n,t}, g_{n,t}\}$, where
           $g_{n,t} = Q_n(o_{n,t}, \phi_{n,t}) + \lambda_{n,t+1}(\tilde{A}_n(t - d_n) - Q_n(t))$;
    if Get a batch of samples then
        Update primal variables to renew $\pi_n$,
            $\boldsymbol{\theta}_{n,t+1} = \boldsymbol{\theta}_{n,t} + \alpha[g_{n,t}\nabla_{\boldsymbol{\theta}} \log \pi_n\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle]$
    end
end
\* Workflow of Edge-based Critic *\
for $t = 1, 2, 3 \ldots$ do
    1. Keep collecting local information,
    if both $s_t, a_t$ and $s_{t+1}, a_{t+1}$ are received then
        generate an experience $e_t$;
    end
    2. Deliver $Q_n(o_{n,t}, \phi_{n,t})$ to each actor $n$ ;
    3. Train the $Q$ network via DQN with $\{e_i\}_{i \in \{1 \ldots \mathcal{E}\}}$
       and the learning rate $\eta$ according to Eqn. (16-17);
end
```

where Q_n' is the target Q network of n and its inputs s_{t+1}, a_{t+1} are also collected from each actor. Here, TD errors indicate whether the performance gets better or worse than expected and is used to adjust both actors and critic in the direction reducing the error mostly. Recall from (17) that Q values consider the future coalition evolution through $Q_n'(s_{t+1}, a_{t+1})$. Based on the feedback of global and long-term Q values, local actors can take the future potential strategies of others into account.

B. Workflow of AoIS

The details of AoIS are summarized in Algorithm 1.

1) Local Actors of D2D Users: The local actor of AoIS aims to learn an age-aware scheduling policy based on primal and dual learning as explained in Section IV-C. To be specific, after the initialization of primal and dual parameters, iterations begin. At the beginning of each time slot t, all the packets generated during the time slot $t - 1$ are stored in the queue. AoIS first updates its age debt by adding up these new packet arrivals and gets the local observation o_n. Then, AoIS makes a scheduling decision on-site through the stochastic policy $\pi_n\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle$. All D2D users form coalitions according to their scheduling actions. As a result, they achieve different data rates and have their packets delivered accordingly. After that, each D2D user computes its age debt $D_{n,t}$ according to (9) and the dual variable $\lambda_{n,t+1}$ is updated so that AoIS can adaptively punish the local actor as in Step 3. Here, P_Λ means the projection into the corresponding feasible set Λ of λ. Since the aim of AoIS is to decrease both AoI violation and power consumption, we set the lower bound of Λ slightly less than 0. Thus, if $D_{n,t} < 0$ and $\lambda_{n,t} < 0$, AoIS will impose some punishment as well. After receiving $Q_n(o_{n,t}, \phi_{n,t})$ from the critic, the D2D user gets a sample $\{o_{n,t}, g_{n,t}\}$ and its primal parameters will be updated according to Step 5 once collecting a batch of samples (with batch size $\mathcal{B}$). Note that this training is performed after each instant of decision making and thus does not need timely feedback from the critic. Compared with a centralized scheduler, AoIS can largely decrease the delay of signaling exchanges between D2D users and the BS since the local actors of AoIS can make decisions instantly and on-site instead of waiting for the orders from a centralized scheduler.

2) Edge-Based Critic: The critic of AoIS is located at the edge controller. It aims to estimate the performance of each actor from a global and long-term view. This can be addressed in the edge with more computation resources. During each time slot t, the critic keeps receiving local observations and data rates of all actors, which form an experience e_t. Then, the edge-based critic learns the Q network, which outputs $Q_n(o_{n,t}, \phi_{n,t})$ for each D2D user and is trained via the techniques of DQN with the learning rate η. To increase data efficiency, off-policy learning (setting both estimated and targeted Q networks) and experience replay can be harnessed to facilitate DQN training. Detailed procedure can be found in [29]. Experience replay method makes the critic accumulate a dataset of experiences and randomly get a minibatch $\{e_i\}_{i \in \{1,\ldots,\mathcal{E}\}}$ of $\mathcal{E}$ samples to train the Q network at each learning trial. By reusing experiences, this can reduce the amount of experiences required for learning, and replace them with more computation and memory in the edge. Note that this training is done in an off-line manner, and thus lags behind the corresponding scheduling actions.

Through this edge-assisted actor-critic learning framework, each D2D user has an intelligent scheduling policy. With this scheduling policy the channels with low channel gain will be ranked low, and the coalition of D2D users (i.e., $\mathbf{C}_k$) will finally converge since the channels with high interference from other D2D users will also be ranked low.

C. Convergence Analysis of AoIS

Since the edge-based critic is trained from the long-term and global perspective, the following theorem can be established for the convergence analysis of AoIS.

Theorem 1: For a system with D2D users following the distributed and local actors of AoIS with a compatible TD(1)[5]

[5] In TD(λ) learning [27], λ is a decay parameter that decides the influence of previous state on the current state.

edge-based critic, if

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty \tag{18}$$

$$\sum_{t=0}^{\infty} \eta_t = \infty, \quad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \quad \lim_{t \to \infty} \frac{\alpha_t}{\eta_t} = 0 \tag{19}$$

the policy gradient of each D2D user n follows

$$\liminf_{t} \mathbb{E}_{\mathbf{H},\mathbf{\Pi}} \left[g_{n,t} \nabla_{\boldsymbol{\theta}} \log \pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle \right] = 0$$

with probability 1.

Proof : According to Lemma 1 and the definition of $g_{n,t}$, the AoIS gradient is given by

$$\begin{aligned} G = {} & \mathbb{E}_{\mathbf{H},\mathbf{\Pi}} \left[\boldsymbol{Q}_n(o_{n,t}, \phi_{n,t}) \nabla_{\boldsymbol{\theta}} \log \pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle \right] \\ & + \mathbb{E}_{\mathbf{H},\mathbf{\Pi}} [\lambda_{n,t+1} (\tilde{A}_n(t - d_n) - Q_n(t)) \nabla_{\boldsymbol{\theta}} \\ & \times \log \pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle]. \end{aligned}$$

First, we consider the expected contribution of the second term to the gradient denoted by G_2 as follows:

$$\begin{aligned} G_2 = {} & \mathbb{E}_{\mathbf{H},\mathbf{\Pi}} [\lambda_{n,t+1} (\tilde{A}_n(t - d_n) - Q_n(t)) \nabla_{\boldsymbol{\theta}} \\ & \times \log \pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle]. \end{aligned}$$

For simplicity, $\pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle$ is rewritten as $\pi_n \langle \phi_{n,t} | o_{n,t} \rangle$. Then, we have

$$\begin{aligned} G_2 = {} & \mathbb{E}_{\mathbf{H}} \Bigg[\sum_{\boldsymbol{\phi}_t^{-n}} \boldsymbol{\pi}_{-n} \langle \boldsymbol{\phi}_t^{-n} | \boldsymbol{o}_t^{-n} \rangle \sum_{\phi_n} \pi_n \langle \phi_{n,t} | o_{n,t} \rangle \nabla_{\boldsymbol{\theta}} \\ & \times \log \pi_n \langle \phi_{n,t} | o_{n,t} \rangle \lambda_{n,t+1} (\tilde{A}_n(t - d_n) - Q_n(t)) \Bigg] \\ = {} & \mathbb{E}_{\mathbf{H}} \Bigg[\sum_{\boldsymbol{\phi}_t^{-n}} \boldsymbol{\pi}_{-n} \langle \boldsymbol{\phi}_t^{-n} | \boldsymbol{o}_t^{-n} \rangle \sum_{\phi_n} \nabla_{\boldsymbol{\theta}} \pi_n \langle \phi_{n,t} | o_{n,t} \rangle \lambda_{n,t+1} \\ & \times (\tilde{A}_n(t - d_n) - Q_n(t)) \Bigg] \\ = {} & \mathbb{E}_{\mathbf{H}} \Bigg[\sum_{\boldsymbol{\phi}_t^{-n}} \boldsymbol{\pi}_{-n} \langle \boldsymbol{\phi}_t^{-n} | \boldsymbol{o}_t^{-n} \rangle \lambda_{n,t+1} (\tilde{A}_n(t - d_n) \\ & - Q_n(t)) \nabla_{\boldsymbol{\theta}} \sum_{\phi_n} \pi_n \langle \phi_{n,t} | o_{n,t} \rangle \Bigg] \\ = {} & \mathbb{E}_{\mathbf{H}} \Bigg[\sum_{\boldsymbol{\phi}_t^{-n}} \boldsymbol{\pi}_{-n} \langle \boldsymbol{\phi}_t^{-n} | \boldsymbol{o}_t^{-n} \rangle \lambda_{n,t+1} (\tilde{A}_n(t - d_n) \\ & - Q_n(t)) \nabla_{\boldsymbol{\theta}} 1 \Bigg] = 0. \end{aligned}$$

This means the second term does not change the expected gradient. Moreover, Q values are learned from a global view, which addresses the issue of nonstationary environment. Thus, the convergence of the first term of expected policy gradient is guaranteed by the traditional actor-critic policy gradient [27]. It proves that a gradient-based actor-critic converges to a local maximum given the following conditions.

1) The policy π is differentiable.

2) $\boldsymbol{Q}$ uses a representation compatible with π.

3) The step size of actor should be negligible compared to the step size of critic so that the actor looks stationary as far as the critic is concerned. Moreover, all learning rates of actor and critic should be nonincreasing so that the learning process could slow down gradually until stop. Hence, the convergence conditions of learning rates should follow (18)-(19). This last condition will guarantee that there is enough time for the critic to evaluate the current actors' policies. Thus, it can be concluded that AoIS converges to the local maximum if the conditions (18)–(19) are satisfied. ■

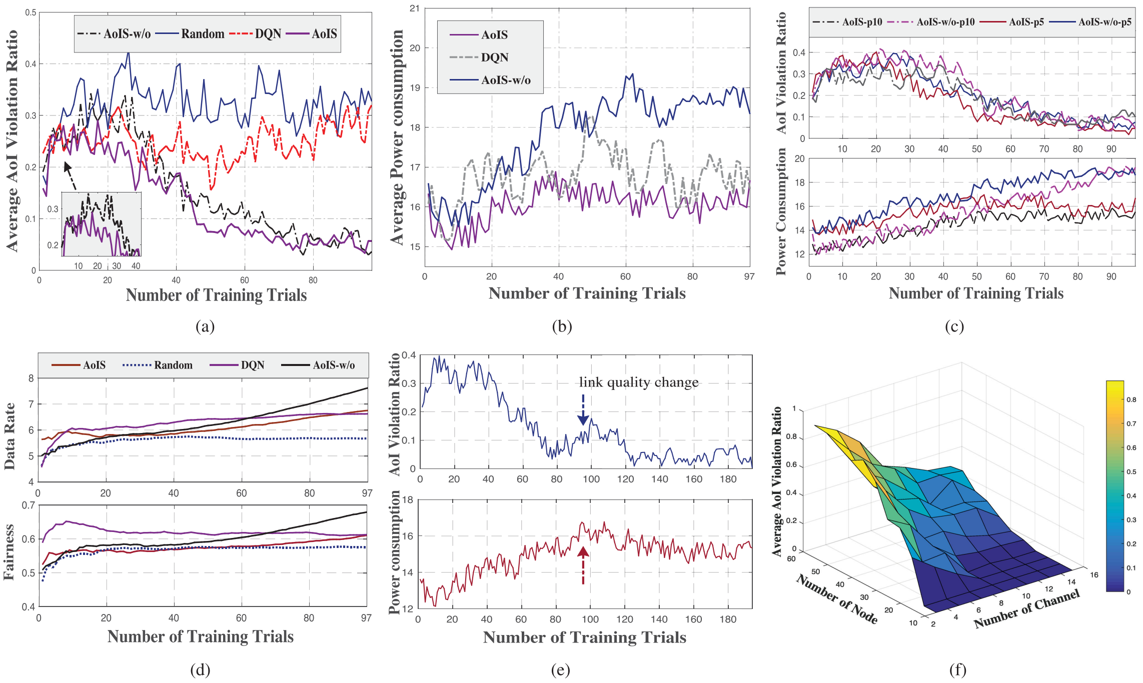
TABLE I
SIMULATION PARAMETERS

Parameter	Value (d[m], f[GHz])
Slot Duration τ	1 ms
RB Bandwidth ω	180 kHz
Packet Size Z	150 Byte
max Tx Power	23 dBm
Noise Power	-174 dBm/Hz
Shadowing st. dev.	12 dB
LOS path loss	$18.7\log(d) + 20\log(f/5) + 46.8$
NLOS path loss	$36.8\log(d) + 20\log(f/5) + 43.8 + 5N_{wall}$

VI. PERFORMANCE EVALUATION

A. Simulation Setup

In this section, the performance of the proposed scheme is evaluated, with the relevant parameters for wireless settings listed in Table I, where d is the distance between two nodes and f is the frequency band that users operate on. Specifically, D2D pairs are randomly positioned inside the 600 m × 600 m industrial service area, where the distance between a pair d is randomly chosen in [10,20]. These D2D devices operate indoors based on the WINNER II Model [30], [31] as well as the simulation settings in [16]. In the Nonline of sight (NLOS) path loss model, N_{wall} denotes the number of walls between two locations and is set to $d/80$. The transmit power is divided into $L = 3$ power levels and the signal to interference plus noise ratio (SINR) threshold SINR_{min} is set to 0 dB. For each source device m, the packet arrival rate and AoI constraint are randomly set between [4,8] and [2,5], respectively. As for the parameters of actors, we have the batch size $\mathcal{B} = 20$, $\Lambda = [-1, 5]$, and $\lambda_0 = 1$. Then, we set the step size $\beta = 0.02$. For $\pi_n \langle \phi_{n,t} | o_{n,t}, \boldsymbol{\theta}_{n,t} \rangle$, we use a three-layer DNN with rectified linear unit (ReLu) activation function, Adam optimizer, initialization Normal$(0, 0.2)$, and initial learning rate $\alpha_0 = 0.002$. As for critic, the structure of $\boldsymbol{Q}$ network starts with an input layer of coalition structure with action embedding. Then, it connects to a fully connected hidden



(a) (b) (c) (d) (e) (f)

Fig. 3. Comparison of AoIS with other scheduling strategies. (a) AoI violation ratio comparison. (b) Power consumption comparison P_{max} = 23 dB. (c) Different power level size comparison. (d) Data rate and fairness comparison. (e) Performance under dynamic link quality. (f) Performance with different *N* and *K*.

layer and output layer with N outputs of Q_n. This is also with ReLu activation function, Adam optimizer, and initialization Normal$(0, 0.2)$. Additionally, the minibatch size $\mathcal{E}$ and the initial learning rate are set to 32 and $\eta_0 = 0.01$, respectively. Both learning rates of actor and critic are multiplied by 0.9 every ten training trials.

B. Simulation Results

The proposed scheme is assessed mainly in terms of the following:

1) average AoI violation ratio, which represents the proportion of the users whose AoI constraints are not satisfied;
2) average data rate of users;
3) average power consumption;
4) Jain's fairness,[6] given by $F(t) = (\sum_{m\in\mathcal{N}} R_m(t))^2/(N\sum_{m\in\mathcal{N}} R_m(t)^2)$.

Based on those metrics, the following scheduling schemes are compared with AoIS:

1) Random, where RB and power levels are randomly selected in each time slot;
2) Decentralized DQN [29], where the DNN (with the same structure as AoIS) outputs the Q value of each action without actor-critic learning;
3) AoIS-w/o, where each D2D user uses policy gradient with its own data rate taken into account and neglects the age debt punishment in reward feedback.

All these schemes are trained locally without the critic.

To compare the performance of various scheduling strategies, Fig. 3(a) first delineates their average AoI violation ratio (N = 15, K = 3). It can be seen that both AoIS and AoIS-w/o outperform other strategies due to the use of policy gradient. They have more stable data rate and lower AoI violation ratio. Obviously, Random has the worst AoI violation performance and the largest violation jitter due to the randomness. Moreover, Decentralized DQN does not function well since it uses the deterministic policy rather than stochastic policy in policy gradient. As a result, it is unable to find the right action due to the fast-varying coalition dynamics and thus the performance is getting worse as time goes by. At last, AoIS performs better than AoIS-w/o since AoIS has a global view based on the edge-based critic as well as more comprehensive rewards (i.e., the dual value $g_{n,t}$ rather than only the data rate $R_n(t)$ as in AoIS-w/o). AoIS-w/o has a relatively low age violation ratio compared with Random and DQN, as it seeks for the action with high data rate and decreases age debt according to (9). However, this is achieved at the expense of high power consumption. It can be seen in Fig. 3(b) that AoIS-w/o achieves the highest power consumption while AoIS has much lower power consumption. Since AoIS considers debt age punishment, λ might decrease below 0 if $D_n(t) < 0$. Hence, AoIS will decrease its power level to not only save energy but also lessen the interference to other coalition members. Moreover, DQN has less power consumption than AoIS-w/o due to inaccurate Q values.

To further analyze the power consumption, Fig. 3(c) compares AoIS and AoIS-w/o in different power level sizes ($L = 5, 10$). It can be observed that AoIS-w/o will consume more energy than AoIS. Both strategies with $L = 10$ have more opportunities to adjust their power levels and thus result in a little less power consumption than that with $L = 5$. However, the action space will extend and thus the convergence rate will be smaller for both AoIS and AoIS-w/o when $L = 10$. Thus, a suitable power

[6] In our case, differentiated data rate (less fairness) is preferable due to users' various age limits and data arrival rates.

level size should be found. Fig. 3(d) shows the average data rate and Jain's fairness of D2D users. Since data rate is the only goal of AoIS-w/o, it receives the highest average data rate of users, however, with the highest power consumption. Although AoIS has relatively lower data rate than AoIS-w/o, DQN, it can achieve the lowest AoI violation ratio. This is because AoIS will make the users with large slack between their age and age limits choose the lower power level by allowing λ to be less than 0. Thus, a low AoI violation ratio can be achieved with just enough data rate for each user rather than the highest. A similar trend can be observed in the second figure in Fig. 3(d). The lower fairness of AoIS compared with DQN and AoIS-w/o denotes that the differentiated data rate supply satisfies the various delivery demands of users.

AoIS is evaluated under dynamic link quality in Fig. 3(e), where the channel fading (according to Rayleigh distribution) is changed at the arrow point. We can see that after AoIS finds out a suitable scheduling policy, AoIS can quickly adjust it according to new environment since AoIS has already learned a satisfactory coalition structure of D2D users. Therefore, the trend of power consumption is relatively stable. Finally, the lowest AoI violation ratio of AoIS is shown in Fig. 3(f) with the increasing number of D2D pairs and RBs. Normally, the performance of AoI violation ratio deteriorates in the crowded networks with fewer resources and more users. It can be seen that the impact of resource size on AoI violation ratio is larger than that of user number since most users can find out suitable coalitions for themselves to improve their data rate in the large-scale networks with enough RBs.

VII. Conclusion

In this article, we proposed an AoI-aware distributed scheduling scheme, namely AoIS, for D2D-enabled IWNs. AoIS guided multiple D2D device pairs to deal with their age constraints and unknown operating environment with primal-dual and actor-critic RL. An edge-based critic could effectively mitigate the nonstationary issue caused by multiple local actors' concurrent learning from a long-term and global view. The problem formulation and optimization process of AoIS provided a theoretical basis for future studies on multiagent constrained learning systems. Moreover, the proposed learning framework of centralized estimation and decentralized execution could further promote the potential of edge-assisted IWNs. For our future work, we will consider the learning-based distributed scheduling for heterogeneous industrial applications in a cooperative setting. For such scenarios, multiple devices with various QoS requirements share a single reward, and thus Q value feedback will be carefully coordinated to improve multiagent learning accuracy.

References

[1] M. Luvisotto, Z. Pang, and D. Dzung, "Ultra high performance wireless control for critical applications: Challenges and directions," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1448–1459, Oct. 2017.

[2] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[3] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," in *Proc. IEEE INFOCOM*, 2016, doi: 10.1109/INFOCOM.2016.7524524.

[4] Z. Zhou, Y. Guo, Y. He, X. Zhao, and W. M. Bazzi, "Access control and resource allocation for M2M communications in industrial automation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3093–3103, Mar. 2019.

[5] Y. Wu, L. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 392–407, Jan. 2019.

[6] L. Liu and W. Yu, "Social-aware incentive mechanisms for D2D resource sharing in IIoT," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5045–5058, Nov. 2018.

[7] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2731–2735.

[8] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Decentralized status update for age-of-information optimization in wireless multiaccess channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2018, pp. 2276–2280.

[9] R. D. Yates and S. K. Kaul, "Status updates over unreliable multiaccess channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2017, doi: 10.1109/ISIT.2017.8006544.

[10] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraint," in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 1844–1852.

[11] I. Kadota, A. Sinha, E. Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE Trans. Netw.*, vol. 26, no. 6, pp. 2637–2650, Oct. 2018.

[12] Q. He, D. Yuan, and A. Ephremides, "Optimal link scheduling for age minimization in wireless systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5381–5394, Aug. 2018.

[13] Y. P. Hsu, E. Modiano, and L. Duan, "Age of information: Design and analysis of optimal scheduling algorithms," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2017, pp. 561–565.

[14] Q. Li, N. Zhang, M. Cheffena, and X. Shen, "Channel-based optimal back-off delay control in delay-constrained industrial WSNs," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 696–711, Oct. 2020.

[15] N. Lyamin, B. Bellalta, and A. Vinel, "Age of information-aware decentralized congestion control in VANETs," *IEEE Commun. Lett.*, vol. 2, no. 1, pp. 33–37, Feb. 2020.

[16] M. K. Abdel-Aziz, S. Samarakoon, C. Liu, M. Bennis, and W. Saad, "Optimized age of information tail for ultra-reliable low-latency communications in vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1911–1924, Dec. 2020.

[17] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1309–1322, Oct. 2015.

[18] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, and M. Hollick, "CBMoS: Combinatorial bandit learning for mode selection and resource allocation in D2D systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2225–2238, Aug. 2019.

[19] B. Lin, X. Wang, W. Yuan, and N. Wu, "A novel OFDM autoencoder featuring CNN-based channel estimation for internet of vessels," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7601–7611, Apr. 2020.

[20] C. Chen, L. Lyu, S. Zhu, and X. Guan, "On-demand transmission for edge-assisted remote control in industrial network systems," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4842–4854, Nov. 2020.

[21] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, Jan. 2018.

[22] H. Wu, F. Lyu, C. Zhou, J. Chen, L. Wang, and X. Shen, "Optimal UAV caching and trajectory in aerial-assisted vehicular networks: A learning-based approach," *IEEE J. Sel. Areas Commun.*, early access, doi: 10.1109/JSAC.2020.3005469.

[23] H. Li, K. Ota, and M. Dong, "ECCN: Orchestration of edge-centric computing and content-centric networking in the 5G radio access network," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 88–93, Jul. 2018.

[24] P. Park, S. C. Ergen, C. Fischione, C. Lu, and K. H. Johansson, "Wireless network design for control systems: A survey," *IEEE Commun. Surv. Tut.*, vol. 20, no. 2, pp. 978–1013, Dec. 2018.

[25] M. Eisen, C. Zhang, L. F. O. Chamon, D. D. Lee, and A. Ribeiro, "Learning optimal resource allocations in wireless systems," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2775–2790, Apr. 2019.

[26] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Sep. 1989.

[27] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003.

[28] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-13, no. 5, pp. 834–846, Sep. 1983.

[29] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[30] "WINNER II path loss model," 2012. [Online] Available: http://www.raymaps.com/index.php/winner-ii-path-loss-model/

[31] Z. Zhu, S. Jin, Y. Yang, H. Hu, and X. Luo, "Time reusing in D2D-enabled cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3185–3200, Feb. 2018.

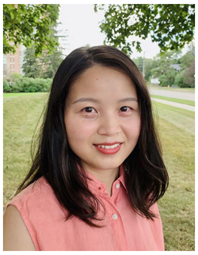
Mingyan Li (Member, IEEE) received the B.Eng. degree in telecommunication engineering from Jinlin University, Changchun, China, in 2015. She is currently working toward the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.

Her current research interests include industrial wireless networks and application in industrial automation, joint design of communication and control in industrial cyber–physical systems, software-defined networking and network slicing, ultrareliable low-latency communication, and mission-critical machine-type communication in industrial wireless networks.

Cailian Chen (Member, IEEE) received the B.Eng. and M.Eng. degrees in automatic control from Yanshan University, Qinhuangdao, China, in 2000 and 2002, respectively, and the Ph.D. degree in control and systems from City University of Hong Kong, Hong Kong, in 2006.

She joined the Department of Automation, Shanghai Jiao Tong University in 2008 as an Associate Professor. She is currently a Full Professor. Before that, she was a postdoctoral research associate with The University of Manchester, Manchester, U.K. (2006–2008). She was a Visiting Professor with the University of Waterloo, Waterloo, ON, Canada (2013–2014). Her research interests include industrial wireless networks and computational intelligence and Internet of Vehicles. She authored three research monographs and more than 100 referred international journal papers. She is the inventor of more than 20 patents.

Dr. Chen is currently an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IET Cyber-Physical Systems: Theory and Applications*, and *Peer-to-Peer Networking and Applications* (Springer). She was a Guest Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the TPC Chair of ISAS'19, the Symposium TPC Co-Chair for IEEE GLOBECOM 2016, the Track Co-Chair for VTC2016-fall and VTC2020-fall, and the Workshop Co-Chair for WiOpt'18. He was the recipient of the prestigious IEEE Transactions on Fuzzy Systems Outstanding Paper Award in 2008, and Best Paper Award of WCSP'17 and YAC'18. She won the Second Prize of National Natural Science Award from the State Council of China in 2018, First Prize of Natural Science Award from The Ministry of Education of China in 2006 and 2016, respectively, and First Prize of Technological Invention of Shanghai Municipal, China in 2017. She was honored Changjiang Young Scholar in 2015 and Excellent Young Researcher by NSF of China in 2016. He has been actively involved in various professional services.



Huaqing Wu (Student Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively. She is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada.

Her current research interests include vehicular networks with emphasis on edge caching, wireless resource management, internet of vehicles, and space–air–ground integrated networks.

Xinping Guan (Fellow, IEEE) received the B.Sc. degree in mathematics from Harbin Normal University, Harbin, China, in 1986, and the Ph.D. degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 1999.

He is currently the Chair Professor of Shanghai Jiao Tong University, Shanghai, China, where he is the Dean of School of Electronic, Information and Electrical Engineering and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. Before that, he was the Executive Director of Office of Research Management, Shanghai Jiao Tong University, a Full Professor, and Dean of Electrical Engineering, Yanshan University, Qinhuangdao, China. As a Principal Investigator, he has finished/been working on more than 20 national key projects. He is the leader of the prestigious Innovative Research Team of the National Natural Science Foundation of China. He is an Executive Committee Member of Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He has authored or coauthored five research monographs, more than 200 papers in IEEE transactions and other peer-reviewed journals, and numerous conference papers. His current research interests include industrial network systems, smart manufacturing, and underwater networks.

Dr. Guan received the Second Prize of the National Natural Science Award of China in both 2008 and 2018, the First Prize of Natural Science Award from the Ministry of Education of China in both 2006 and 2016. He was a recipient of the IEEE Transactions on Fuzzy Systems Outstanding Paper Award in 2008. He is a National Outstanding Youth honored by NSF of China, and Changjiang Scholar's by the Ministry of Education of China and State-Level Scholar of New Century Bai Qianwan Talent Program of China.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular ad hoc and sensor networks.

Dr. Shen was/is the Editor-in-Chief for the IEEE INTERNET OF THINGS journal, IEEE NETWORK, *IET Communications*, and *Peer-to-Peer Networking and Applications*. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society. He was the Technical Program Committee Chair/Co-Chair for the IEEE Globecom'16, the IEEE Infocom'14, the IEEE VTC'10 Fall, the IEEE Globecom'07, the Symposia Chair for the IEEE ICC'10, and the Chair for the IEEE Communications Society Technical Committee on Wireless Communications. He is the elected IEEE Communications Society Vice-President for Technical and Educational Activities, Vice-President for Publications, Member-at-Large on the Board of Governors, Chair of the Distinguished Lecturer Selection Committee, and Member of IEEE Fellow Selection Committee. He was the recipient of the R.A. Fessenden Award in 2019 from the IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) presented in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society, and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He was also the recipient of the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premiers Research Excellence Award in 2003 from the Province of Ontario, Canada.