AI-aided Traffic Control Scheme for M2M Communications in the Internet of Vehicles

Haijun Zhang†, Minghui Jiang†, Xiangnan Liu†, Keping Long†, and Victor C. M. Leung‡
† University of Science and Technology Beijing, Beijing, China,
Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services.
‡ The University of British Columbia, Vancouver, Canada.

Abstract—Due to the rapid growth of data transmissions in internet of vehicles (IoV), finding schemes that can effectively alleviate access congestion has become an important issue. Recently, many traffic control schemes have been studied. Nevertheless, the dynamics of traffic and the heterogeneous requirements of different IoV applications are not considered in most existing studies, which is significant for the random access resource allocation. In this paper, we consider a hybrid traffic control scheme and use proximal policy optimization (PPO) method to tackle it. Firstly, IoV devices are divided into various classes based on delay characteristics. The target of maximizing the successful transmission of packets with the success rate constraint is established. Then, the optimization objective is transformed into a markov decision process (MDP) model. Finally, the access class barring (ACB) factors are obtained based on the PPO method to maximize the number of successful access devices. The performance of the proposal algorithm in respect of successful events and delay compared to existing schemes is verified by simulations.

I. INTRODUCTION

By the quick development and widespread use of the mobile internet, the internet of vehicles (IoV) achieves the sharing of information between vehicles through advanced wireless communication technology [1]. It provides compelling services for intelligent traffic system (ITS) such as intelligent warning, real-time road condition monitoring and traffic management [2]. To realize large-scale and high-volume data interaction, numerous communication units and sensors are used in IoV [3]. With the exponential growth of on-road devices connected to the internet, the IoV networks will suffer from traffic congestion problem and poor quality of service [4], especially for emergency services, which requires ultra-reliable low delay to ensure safe driving [5]. Therefore, how to obtain low delay with the growing number of successful devices access is becoming a big issue in the face of the proliferation of devices [6].

Access class barring (ACB) mechanism is viewed as a prospective technique due to its effectiveness in controlling bursts of traffic. In [7], a random access preambles allocation problem based on the dynamic ACB (DACB) algorithm was investigated. In [8], the authors allocated random access channel (RACH) resources and retransmission metrics to various devices as required to address overloaded problem. Two classes of devices were proposed in [9], the devices that sensitive to delay were classified as high priority. However, most studies are based on the assumption that the base station (BS) is completely known about the results in last random access opportunity (RAO), whereas the BS only knows the random access results of the amount of successful accesses, collisions, and unselected preambles actually.

Deep reinforcement learning (DRL) [10] is an efficient machine learning method, which can achieve effective traffic control based on the information known to the BS, there is no necessity to estimate the number of backlog nodes. A DRL multiple access protocol was proposed to maximize throughput and achieve fairness in [11]. In [12], a deep Q network (DQN)-based resource scheduling method for the traffic control was proposed, where the total transmission rate of uplink was maximized.

Although some works have been done for DRL-based
dynamic traffic control, these papers mainly focus on the one class of devices. However, it is unreasonable to apply a uniform ACB factor to all devices, due to the heterogeneous QoS requirements of various IoV applications. However, as an emerging DRL algorithm, the proximal policy optimization (PPO) [13] based dynamic ACB scheme for multiple classes in IoV networks has not been investigated well in the current works.

In this paper, we investigate the traffic control problem based on priority-based dynamic access class barring (PDACB) scheme and a novel learning strategy with consideration of success rate requirements and back-off (BO) scheme in IoV networks. The purpose is to maximize the successful transmission of packets while reducing the delay for the high priority class.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

In this paper, we consider the machine to machine (M2M) communications in IoV under the device-dense situation as shown in Figure 1, where IoV is composed of a BS, a certain number of devices comprising on-board units (OBU) and sensors. The devices are divided into 2 classes of priority depending on the delay sensitivity of each device. OBUs used for data communication between vehicles and roads with characteristics of low volume and delay sensitivity are classified as high priority users. Sensors with characteristics of high volume and delay tolerance are classified into lower priority users. Devices of each classes randomly generate uplink packets, the massive devices send packets with a period of TA. The probability distribution of packet generation is given by

\[ g(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1} B(\alpha, \beta)}, \]

where \( B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \), \( \alpha = 3 \), \( \beta = 4 \), TA is the activation period. The new packets generated at \( t \)-th slot is formulated by

\[ \lambda_k(t) = N_k \int_{t_{k-1}}^{t_k} g(t) dt, \]

where \( N_k \) is the amount of devices for class \( k \).

In this paper, a hybrid scheme combining ACB scheme and BO scheme is considered to alleviate the large-scale devices access problem before RACH. Specifically, the ACB scheme is started at the beginning of each RAO, the BS generates an ACB factor \( p \), the device attempting to access the IoV networks chooses a number \( q \). If \( q \leq p \), the device passes the ACB check. If not, the device needs to retry and restart the next attempt. The waiting time required until the next attempt can be expressed as \( T_{barring} = (0.7 + 0.6u) T_{ACB} \).

where \( u \in \mathcal{U}[0, 1], \mathcal{U} \) denotes uniform distribution. \( T_{ACB} \) indicates a barring time. In other words, it needs to wait for \( l \in \mathcal{U}[l_{\text{min}}, l_{\text{max}}] \) slots, where \( l_{\text{min}} = \frac{0.7 T_{ACB}}{T_s}, l_{\text{max}} = \frac{1.3 T_{ACB}}{T_s} \).

The random access process is carried out after ACB checking. In a RACH, the device needs to transmit a set of codes called preamble to obtain the uplink channel. There are 64 orthogonal preambles in each RAO, which are shared by all devices attempting access. The procedure of four handshakes in competition-based mode is summarized as follows [14], [15]. Firstly, the device trying to transmit packets chooses a preamble and sends it to the BS. Secondly, the BS calculates an identifier (ID) for each preamble and sends a random access response message to the device. Thirdly, the device sends a connection request message to the BS. The devices using the same preamble will send their packets in the same time-frequency RB, causing the BS to be unable to decode the received connection request message, thus a conflict occurs. Finally, the BS sends a contention resolution message to the device in response, and the device verifies whether the access is successful by checking the ID contained in the contention resolution message. The required waiting time for collision devices to retry can be expressed as \( T_{BO} = \mathcal{U}[0, T_b] \), where the \( T_b \) is BO parameter. That is, the time slots for retrying is \( b \in \mathcal{U}[1, B] \), where \( B = \frac{T_s}{T_s} \).

\( T_s \) is the time interval between RAOs, i.e. time slot.

B. Problem Formulation

We analyze the traditional single class ACB algorithm firstly. There are \( N \) devices in the IoV networks that periodically generate packets and try to connect to the BS. The expected number of ACB check that can be successfully passed is denoted as \( N_p = N \ast p \).

The amount of devices for which failed in ACB check can be formulated as \( N_f = N - N_p \). The device passed the ACB check chooses a preamble among the \( F \) preambles. Successful access is finished when the preamble is chosen by only one device. Let \( N_s \) denote the number of devices successfully accessed, the expected number of devices that successfully accessed can be written as

\[ \mathbb{E}[N_s | N_p] = F \left( \frac{N_p}{1} \right) \frac{1}{F} \left( 1 - \frac{1}{F} \right)^{N_p-1} \]

\[ = N_p \left( 1 - \frac{1}{F} \right)^{N_p-1}. \]

The differentiation of (3) can be calculated by

\[ \frac{d\mathbb{E}[N_s | N_p]}{dN_p} = (1 - \frac{1}{F})^{N_p-1} + N_p (1 - \frac{1}{F})^{N_p-1} \ln \left( 1 - \frac{1}{F} \right) = 0. \]
According to the approximation $\ln \left(1 - \frac{1}{p}\right) \simeq -\frac{1}{p}$ and $-\frac{1}{p} \simeq 0$, the solution to (4) is given by

$$N_p^* = -\ln \left(1 - \frac{1}{p}\right)^{-1} \simeq -\left(\frac{1}{p}\right)^{-1} = F.$$  \hspace{1cm} (5)

Therefore, the number of devices successfully accessed is maximized when $N_p^*$ is equal to $F$. According to $N_p = N^* - p$, the successful access maximization problem can transform into the problem of finding the optimal ACB factor.

Compared to the traditional ACB algorithm, the ACB with $K$ classes is considered. The $K$ ACB factors are adjusted dynamically in each time slot. The expected number of successfully accessed for class $k$ is written as

$$N_{s,k} = \sum_{i} N_{s(k)} = \sum_{i} \frac{n(i)-1}{1-F}$$

where $N_{s(k)} = N_{s} + k$ is the amount of class $k$ attempting transmit packet and $n(i)$ is the amount of devices sending access requests for all classes.

Since devices that failed in the ACB check or in conflict need to wait a random time before retrying to access, the amount of devices attempting transmit packets for the $i$-th slot is given by

$$n_{k}(i) = \begin{cases} n_{k}(i) + \frac{n(i)-1}{1-F} & i = 1 \\ \frac{n(i)-1}{1-F} & 1 < i \leq \frac{T_{k}}{T_{k}} \\ \frac{n(i)-1}{1-F} & i > \frac{T_{k}}{T_{k}} \end{cases}$$

where $L = l_{max} - l_{min} + 1$, the amount of collisions in class $k$ is written as $N_{c,k} = N_{s,k} - N_{s,k}^*$.

The devices that failed in the ACB check or in conflict need to wait a random time before retrying to access. The maximum of successful device accesses can be obtained while satisfying various successful access rate constraints

$$\max \sum_{i=1}^{l_{max}} n_{i}(i) \left(1 - \frac{1}{p}\right)^{-1} s.t. \begin{cases} N_{s,k} \geq \beta_k, \\ 0 \leq p_k \leq 1 \end{cases}$$

where $\beta_k$ is the ratio of successful accessed that we expect for class $k$.

III. DEEP REINFORCEMENT LEARNING BASED RACH OPTIMIZATION

A. Proximal Policy Optimization

Compared with the trust region policy optimization, PPO introduces clipped surrogate objective instead of constraint function, which reduces the complexity of the algorithm.

1) Clipped Surrogate Objective: The clipped surrogate objective in PPO algorithm is formulated by

$$J_{CLIP}^\theta (\theta) = \mathbb{E}_t \left[ \min (\sigma_t(\theta), clip(\sigma_t(\theta), 1 - \mu, 1 + \mu)) A_t(\theta_{old}) \right],$$

where $\sigma_t(\theta) = \pi_\theta (a_t | s_t) / \pi_{old} (a_t | s_t)$ is the ratio of the new policy to the old policy, and $A_t(\theta_{old}) = A_{\pi_{old}} (s_t, a_t)$. To avoid large gaps between new policy $\pi_\theta (a_t | s_t)$ and old policy $\pi_{old} (a_t | s_t)$, the PPO uses clipped probability ratio to limit $r_t(\theta)$ to $[1 - \mu, 1 + \mu]$.

2) Generalized Advantage function: Advantage function is an important component of the PPO for the stability of the algorithm, which can be formulated as

$$A_t(\theta_{old}) = -V_{\pi_{old}} (s_t) + r_t + \cdots + \gamma_{T-t+1}^t r_{T-t+1} + \gamma_{T-t}^t V_{\pi_{old}} (s_T).$$

Since the state-value function $V (s)$ is used in (10), the objective of PPO algorithm combines the clipped surrogate objective $J_{CLIP}^\theta (\theta)$ and the value function error term $L_t^\gamma (\theta)$. Furthermore, an entropy bonus $E_{\pi_{old}} (s_t)$ is added to allow for full exploration. Thus, the objective of PPO algorithm is written by

$$J_t^{PPO} (\theta) = \mathbb{E}_t \left[ J_{CLIP}^\theta (\theta) - c_1 L_t^\gamma (\theta) + c_2 E_{\pi_{old}} (s_t) \right],$$

where $c_1$ and $c_2$ are coefficients, $L_t^\gamma (\theta) = (V_{\pi_{old}} (s_t) - V_{t_{\text{arg}}})^2$.

B. PPO-based RACH Optimization

The problem of traffic control is transformed into a DRL problem, which mainly includes the design of action space, state space and reward.

1) Action Space: The DRL agent is deployed at the BS to observe the access results in last slot and output action according to the policy $\pi (a_t | s_t, \theta)$ after a multi-step exploration and learning process. The priority is classified according to delay requirement of each device, the action space can be expressed as

$$a_t = \left\{ p_{1(t)}, p_{2(t)}, \cdots, p_{K(t)} \right\}. $$

2) State Space: The information known to the BS is the number of devices successfully accessed $N_{s(t-1)}$, collision $N_{c(t-1)}$ and the unused preambles $N_{idle(t-1)}$, considering the results of the analysis from (5), maximum of successful access is achieved when $N_{s(t-1)}$ is approximately equal to $F$. Therefore, $N_{dist(t-1)} = N_{s(t-1)} - F$ is considered as one of the features. The state of the environment observed by the DRL agent in step $t$ is

$$s_t = \left\{ N_{s(t-1)}, N_{dist(t-1)}, N_{idle(t-1)} \right\}. $$
where $N_{s}^{(t-1)} = (N_{s,1}^{(t-1)}, N_{s,2}^{(t-1)}, \ldots, N_{s,K}^{(t-1)})$, $\mathbf{N}_{\text{dist}}^{(t-1)} = (N_{\text{dist,1}}^{(t-1)}, N_{\text{dist,2}}^{(t-1)}, \ldots, N_{\text{dist,K}}^{(t-1)})$.

**Algorithm 1** PPO-based Random Access Optimization in IoV Networks

**Initialize:** $\theta$ is initialized by using random numbers generated by uniform distribution and empty the buffer $B$.

for *Iteration* $t = 1$ to $T$ do

Obtain initial observation state $s_1$ and buffer $B$;

for $t = 1$ to $T$ do

Choose action $a_t$ according to the old policy $\pi(s_t | \theta_{old})$;

Execute action $a_t$ and receive next state $s_{t+1}$, reward $r_t$;

Collect transition $(s_t, a_t, r_t, s_{t+1})$ to buffer $B$;

Calculate advantage estimates $A_1, \cdots, A_T$ according to (10);

end for

for *epoch* $e = 1$ to $M$ do

Sample a random minibatch of $N$ transitions $(s, a, r, s)$ from buffer $B$;

Calculate $J^{\text{PPO}}(\theta)$ according to (11);

Calculate $\nabla_\theta J^{\text{PPO}}(\theta)$ by gradient method;

Update $\theta$ according to $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

$\theta_{old} \leftarrow \theta$;

end for

end for

3) **Reward:** The design of the reward is crucial in DRL, which directly affects the convergence speed and learning effect of the agent. The reward for class $k$ in step $t$ is defined as

$$r_k^{(t)} = N_{s,k}^{(t)} - \alpha_k \left| N_{p,k}^{(t)} - F \right|,$$

where $\alpha_k$ is coefficient, the total reward can be written as

$$r_t = \sum_{k=1}^{K} \beta_k r_k^{(t)}.$$  \hspace{1cm} (15)

**IV. SIMULATION RESULTS AND DISCUSSIONS**

In this section, the traffic control scheme with multiple ACB factors based on the DRL algorithm is verified by simulation results. The period $T_A = 20$ s, the time slot $T_s = 20$ ms, the number of contention-based preambles $F = 54$. The barring time $T_{ACB}$ and BO parameter value $T_b$ are 0.2 s and 80 ms, respectively.

Fig. 2 shows the reward of PPO-based PDACB scheme versus the number of iterations. The rapid convergence of PPO-based algorithm by using only 50 iterations is achieved in the Fig. 2, which confirms the excellent performance of the state spaces and rewards designed in this paper.

Fig. 3 shows the amount of successfully accessed IoV devices comparison of PDACB scheme and DACB with class 1 and class 2 versus the number of iterations. It is can be seen that the PDACB scheme can satisfy the heterogeneous QoS requirements, and the PDACB scheme is better than the existing DACB scheme in respect of maximizing the amount of successfully accessed devices.

Fig. 4 shows access delay with class 1 and class 2 of proposed scheme versus the number of iterations. In the PDACB scheme, it is obvious that class 1 has lower delay than class 2. The delay of class 1 in proposed scheme is 66.7% lower than DACB scheme.

Fig. 5 shows the random access efficiency with class 1 and class 2 versus iterations. Random access efficiency is denote the proportion of successfully transmitted packets...
to all packets. The weighted coefficient $\beta_1$ and $\beta_2$ in (15) are 0.8 and 0.2. Obviously, the random access efficiency of PDACB scheme is higher than the DACB scheme’s.

V. CONCLUSIONS

In this paper, a priority based hybrid traffic control scheme with PPO algorithm for massive IoV M2M communication has been proposed. To satisfy the delay requirements for various devices, devices have been divided into different classes. Multiple ACB factors change dynamically to adapt the traffic in time, and the maximum of successful accesses devices under the successful access rate constraint. PPO algorithm has been utilized to seek the optimal ACB factors. Simulation results show that PDACB scheme outperform DACB scheme in respect of the amount of successfully transmitted packets and delay.

ACKNOWLEDGMENT

This work is supported in part by National Key R&D Program of China (Grant No. 2019YFB1803304), in part by Beijing Natural Science Foundation (L212004), in part by the State Key Laboratory of Advanced Metallurgy under Grant KF20-04, in part by the Fundamental Research Funds for the Central Universities under Grants FRFTP-19-002C1 and RC1631, and in part by Beijing Top Discipline for Artificial Intelligent Science and Engineering, University of Science and Technology Beijing. The corresponding authors are Haijun Zhang and Keping Long.

REFERENCES

[1] S. Chen et al., “Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G,” IEEE Commun. Stand. Mag., vol. 1, no. 2, pp. 70–76, Jun. 2017.
[2] A. Haydari and Y. Yilmaz, “Deep reinforcement learning for intelligent transportation systems: A survey,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 1, pp. 11–32, Jan. 2022.
[3] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, “Challenges and solutions for cellular based V2X communications,” IEEE Commun. Surv. Tutor., vol. 23, no. 1, pp. 222–255, Firstquarter 2021.
[4] Y. Liu, M. Xiao, S. Chen, F. Bai, J. Pan, and D. Zhang, “An intelligent edge-chain-enabled access control mechanism for IoV,” IEEE Internet Things J., vol. 8, no. 15, pp. 12231–12241, Aug. 2021.
[5] H. S. Jang, H. Jin, B. C. Jung, and T. Q. S. Quek, “Versatile access control for massive IoT: Throughput, latency, and energy efficiency,” IEEE Trans. Molt. Comput., vol. 19, no. 8, pp. 1984–1997, Aug. 2020.
[6] N. Xia, H.-H. Chen, and C.-S. Yang, “Radio resource management in machine-to-machine communications—a survey,” IEEE Commun. Surv. Tutor., vol. 20, no. 1, pp. 791–828, Firstquarter 2018.
[7] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, “D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks,” IEEE Trans. Veh. Technol., vol. 65, no. 12, pp. 9847–9861, Dec. 2016.
[8] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, “PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks,” IEEE Trans. Veh. Technol., vol. 63, no. 5, pp. 2467–2474, Jun. 2014.
[9] J. Liu, M. Agiwal, M. Qu, and H. Jin, “Online control of preamble groups with priority in massive IoT networks,” IEEE J. Sel. Areas Commun., vol. 39, no. 3, pp. 700–713, Mar. 2021.
[10] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[11] Y. Yu, T. Wang, and S. C. Liew, “Deep-reinforcement learning multiple access for heterogeneous wireless networks,” IEEE J. Sel. Areas Commun., vol. 37, no. 6, pp. 1277–1290, Jun. 2019.
[12] M. Chu, H. Li, X. Liao, and S. Cui, “Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems,” IEEE Internet Things J., vol. 6, no. 2, pp. 2009–2020, Apr. 2019.
[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017. [Online]. Available: arXiv:170706347.
[14] “Radio resource control (RRC); protocol specification,” Tech. Rep. 3GPP TS36.331 v15.1.0, Apr. 2018.
[15] “Medium access control (MAC) protocol specification,” Tech. Rep. 3GPP TS36.321 v15.1.0, Mar. 2018.