2024 27th International Symposium on Wireless Personal Multimedia Communications (WPMC) | 979-8-3503-9231-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/WPMC63271.2024.10863142

Reinforcement Learning-Based Optimization of Relay Selection and Transmission Scheduling for UAV-Aided mmWave Vehicular Networks

Aditya Guhagarkar*, Thushan Sivalingam†, Vimal Bhatia*†, Nandana Rajatheva†, and Matti Latva-aho†
*Department of Electrical Engineering, IIT Indore, 453552, India
†Centre for Wireless Communications, University of Oulu, Oulu, Finland
Email: ee210002005@iiti.ac.in

***Abstract*—Millimeter-wave (mmWave) communications offer abundant bandwidth for vehicular networks, however it is prone to blockages due to buildings, topology and other environmental factors. To address these challenges, we propose a novel unmanned aerial vehicle (UAV)-aided two-way relaying system to enhance vehicular connectivity and coverage. We formulate a joint optimization problem for relay selection and transmission scheduling to minimize transmission time while ensuring throughput requirements. Proximal policy optimization, deep Q-network, and constraint programming models are employed to solve the optimization problem. Extensive evaluations reveal that the proximal policy optimization model achieves close to 100% accuracy.**

***Index Terms*—Concurrent scheduling, deep Q-network, proximal policy optimization, relay selection, vehicular networks.**

I. Introduction

THE rising data demands in vehicular networks necessitate using millimeter-wave (mmWave) bands for next-generation communication. However, mmWave vehicle-to-vehicle (V2V) communications face high path loss and frequent blockages, thereby degrading link quality. Incorporating relay-aided transmission with unmanned aerial vehicle (UAV)s and terrestrial vehicles can enhance coverage, throughput, and reliability, since the UAVs offer flexible deployment and improved connectivity in dynamic environments [1].

Optimizing UAV resources and scheduling is challenging in these settings due to environmental, relative mobility, power, payload and other limitations. Traditional methods struggle with the unpredictability of vehicular networks. Reinforcement learning (RL) techniques, including deep Q-networks (DQN) and proximal policy optimization (PPO), provide solutions by learning optimal scheduling strategies through interaction and feedback [2]. DQN models estimate long-term rewards for scheduling actions, while PPO models optimize policies in continuous action spaces, crucial for managing mmWave V2V communications.

A. Related Works

Relay-aided mmWave communications have shown potential in improving system performance. Wu et al. [3] demonstrated that two-hop device-to-device relaying enhances coverage and spectral efficiency in mmWave networks. Ruiz et al. [4] proposed optimal relay positioning for 5G. UAVs as flexible relays have further improved vehicular communication links [1], with Jing et al. [5] optimizing UAV-assisted systems through joint resource allocation and scheduling to reduce latency and maximize throughput in dynamic environments. The paper introduced the joint relay selection with dynamic scheduling (JRDS) scheme. Efficient scheduling is key to optimizing mmWave communication. Hadded et al. [6] highlighted the effectiveness of time-division multiple access (TDMA) for vehicular applications. Qiao et al. [7] introduced a multi-hop concurrent transmission scheme leveraging the spatial capacity of mmWave relay systems, outperforming traditional single-hop methods. Despite these advancements, challenges in adapting to real-world scenarios persist.

B. Motivations and Contributions

Reliability remains as one of the major challenges for mmWave vehicular communications when terrestrial relays may fail. Terrestrial relays may fail in V2V links among distant transceivers, and UAV relays are limited by battery life under high traffic. However, by using two-way relaying, we may be able to overcome this. Thus in this work, we consider two-way relaying with UAVs. We have developed and evaluated three models: DQN, PPO, and constraint programming, comparing them against a traditional TDMA algorithm and the JRDS scheme to assess their effectiveness in optimizing network performance and reliability in mmWave vehicular communications.

Rest of the paper is organized as follows: Section II introduces the system model. Section III formulates the joint scheduling problem. Section IV details the data processing and proposed schemes. Section V presents simulation results. Finally, Section VI concludes the paper and discusses future research directions.

II. System Overview

This section presents an overview of the mobility models, antenna patterns, and channel models utilized in the proposed UAV-aided mmWave vehicular network.

Part of this work was supported by the FICORE project, and the Nokia Foundation.

A. Mobility Models

1) Vehicle Mobility Model: We consider a snapshot within a time frame where vehicles are aligned along a road, maintaining a minimum safe distance of two meters to avoid collisions. Each vehicle is equipped for V2V communication within the network.

2) UAV Mobility Model: Multiple UAVs are deployed to relay communications between vehicles. Each UAV hovers at a fixed altitude h_u with a 500 m non-overlapping coverage radius, ensuring optimal relay services to enhance network connectivity.

B. Antenna Pattern

Vehicles operate in full-duplex (FD) mode, allowing simultaneous transmission and reception, while UAVs use half-duplex (HD) mode, alternating between transmitting and receiving. The maximum antenna gain is denoted as G_0.

C. Channel Models

1) V2V Links: The channel power gain G for V2V links follows a Gamma distribution with parameter m [8]. The received signal power at receiver r_i from transmitter s_i is expressed [9] as

$$P_r(s_i, r_i) = k_v P_t G_0 g(s_i, r_i) d_{s_i,r_i}^{-\alpha_v}, \tag{1}$$

where k_v is a constant, P_t is the transmit power, d_{s_i,r_i} is the distance between s_i and r_i, and α_v is the path loss exponent. To enhance network performance, vehicles use zero-forcing beamforming (ZFBF) [10] to suppress mutual interference during concurrent transmissions. The SNR at r_i is given by

$$\text{SNR}_{s_i,r_i} = \frac{P_r(s_i, r_i)}{N_0 W}, \tag{2}$$

where N_0 is the noise power spectral density and W is the bandwidth. The data rate R_{s_i,r_i} is determined by the Shannon capacity [11] as

$$R_{s_i,r_i} = \eta W \log\left(1 + \text{SNR}_{s_i,r_i}\right), \tag{3}$$

with η as the transceiver efficiency factor.

2) U2V Links: UAVs provide relay services from an altitude h_u, establishing LoS links with vehicles. U2V links experience large-scale path loss and small-scale Rician fading. The SNR from UAV u to vehicle r_k is expressed as

$$\text{SNR}_{u,r_k} = \frac{k_u P_u G_0 d_{u,r_k}^{-\alpha_u}}{\Omega_{u,r_k} N_0 W}, \tag{4}$$

where k_u is a constant, P_u is the UAV transmit power, and d_{u,r_k} is the distance between UAV u and vehicle r_k. The small-scale fading Ω_{u,r_k} follows a non-central χ^2-distribution [12] as

$$\begin{aligned} f_{\Omega_{u,r_k}}(\omega) = {} & \frac{(K+1)}{\Omega_{u,r_k}} \exp\left(-\frac{K+1}{\Omega_{u,r_k}}\omega\right) \\ & \times I_0\left(\sqrt{\frac{2K(K+1)\omega}{\Omega_{u,r_k}}}\right), \end{aligned} \tag{5}$$

where K is the Rician factor, and $I_0(\cdot)$ is the zero-order modified Bessel function of the first kind.

D. Dynamic Scheduling

Dynamic scheduling (DS) optimizes communication by adapting to network conditions through two phases: scheduling and transmission (see Fig. 1). In the scheduling phase, time slots and channels are allocated based on current conditions and traffic demands, minimizing conflicts. During transmission, data is sent over the allocated slots, enabling simultaneous non-conflicting transmissions, thus improving throughput. DS factors in quality of service and channel conditions to adapt to traffic changes, optimizing resource allocation and reducing idle times, which enhances system performance and reliability. Fig. 1 illustrates this with five flows: non-adjacent flows (e.g., T1, T2) share time slots, while adjacent flows (e.g., T1, T4) require separate slots.

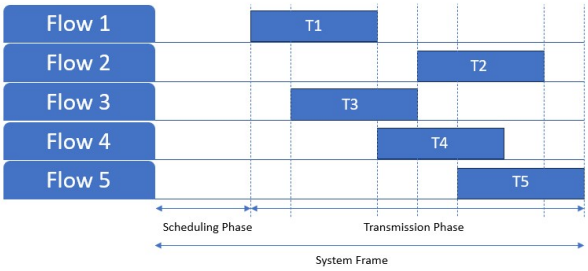

Fig. 1: Illustration of dynamic scheduling.

III. Problem Formulation

Consider a two-lane highway with vehicles equipped with full-duplex antennas. UAVs with specific coverage radii assist in communication. The objective is to schedule n data flows to minimize the required time slots while maintaining throughput requirements. Adjacent flows, sharing a transmitter or receiver, cannot transmit simultaneously unless an intermediate hop is used. Each data flow must meet or exceed a predefined throughput threshold.

A. Constraints

The constraints for this problem are defined as follows:

1) Adjacent Flows Constraint: Data flows that are adjacent, sharing either a transmitter or receiver, must not be scheduled for simultaneous transmission. Mathematically, if flows f_i and f_j are adjacent, then:

$$\text{Transmission}(f_i) \cap \text{Transmission}(f_j) = \emptyset. \tag{6}$$

2) Hops Constraint: The transmission between a pair of nodes is limited to at most two hops. Let $H_{i,j}$ denote the number of hops between transmitter i and receiver j. Then:

$$H_{i,j} \leq 2. \tag{7}$$

3) Throughput Threshold: Each data flow f_i must achieve a network throughput Q_i that meets or exceeds a specified threshold Q_{min}. Thus:

$$Q_i \geq Q_{\text{min}}. \tag{8}$$

4) Mobility Model Constraint: The mobility models of vehicles and UAVs must be considered, as defined in the simulation model.

B. Objective

The goal is to minimize the total number of time slots required for transmitting all data flows.

IV. Proposed Approach

We compare two reinforcement learning models, DQN and PPO, with a constraint programming-based solution, a TDMA algorithm and the JRDS scheme.

A. Data Processing

Data processing involves acquiring vehicle positions and scheduling flows. Each flow's time is calculated by

$$\xi_a = \frac{Q_a \cdot M \cdot T}{R_a \cdot T},$$

where ξ_a is the number of time slots, Q_a is the flow throughput (Gbps), M is a multiplier, T is the slot time, and R_a is the flow data rate.

Flows are filtered by throughput and checked for LoS. Flows with LoS are retained; others are routed via cars or UAVs. If both routes are feasible, the one with the shortest completion time is chosen.

B. Reinforcement Learning Models

We use DQN and PPO due to their adaptability to dynamic environments and scalability for large problems.

1) DQN: DQN approximates Q-values with a neural network, representing the expected reward of actions. It uses experience replay and target networks for stability and improved performance.

2) PPO: PPO trains a policy by optimizing parameters to maximize expected cumulative rewards, ensuring stable updates within a specified range.

Fig. 2 shows the flow graphs for both models, illustrating their architectures and training processes.

C. Constraint Programming Model

This section describes a constraint programming model using the OR-Tools library for efficient network flow scheduling, aimed at minimizing total transmission time.

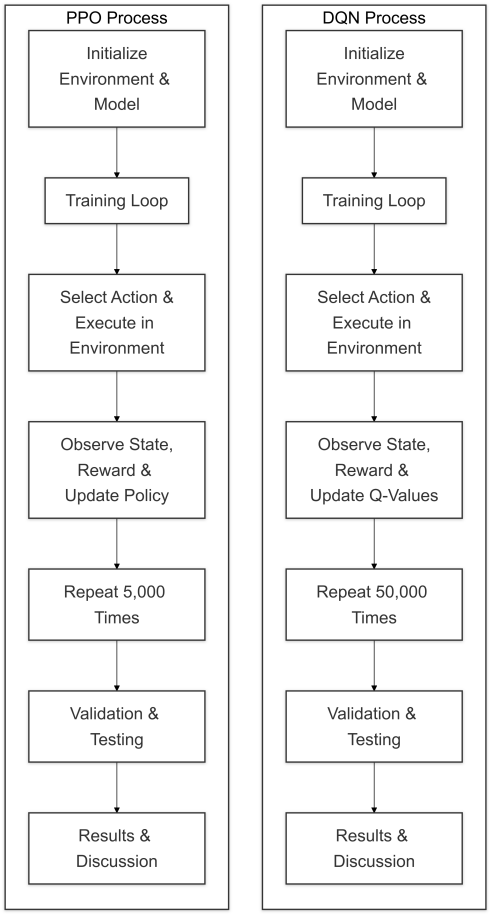

Fig. 2: Flow graphs for DQN and PPO models.

Algorithm 1 Flow Scheduling using Constraint Programming

Input: Set of flows $F = \{f_1, f_2, \ldots, f_n\}$
Output: Optimized schedule
Step 1: Adjacency Matrix Creation
for each pair of flows (f_i, f_j) **do**
  if f_i and f_j share the same transmitter or receiver **then**
    Mark flows f_i and f_j as adjacent in the matrix
  end if
end for
Step 2: Define Variables
Define $start\_times[i]$ for each flow $f_i \in F$
Define $total\_time$ as the total number of time slots used
Step 3: Apply Constraints
for each pair of adjacent flows (f_i, f_j) **do**
  Enforce non-overlapping schedule:
    $start\_times[i] + duration(f_i) \leq start\_times[j]$ **OR**
    $start\_times[j] + duration(f_j) \leq start\_times[i]$
end for
$total\_time \geq \max_{f_i \in F} (start\_times[i] + duration(f_i))$
Step 4: Objective Function
Minimize $total\_time$
Step 5: Solve the Model
Use OR-Tools CP-SAT solver to find the optimal schedule

D. Throughput Calculations

Throughput is defined as the ratio of total data in the given time to total time:

$$\text{Throughput} = \frac{\text{Total Data}}{\text{Total Time}}. \tag{9}$$

Total data is the summation of $R_{s_i,r_i} \times T_i$ for $i = 1$ to N:

$$\text{Total Data} = \sum_{i=1}^{N} R_{s_i,r_i} \times T_i. \tag{10}$$

The total time is $T_{\text{total}} = n \times 0.1$, where n is the total number of time slots, T_i is the time taken for flow i, and 0.1 is the duration of each time slot in seconds. Substituting these into the throughput equation gives:

$$\text{Throughput}\,(\mathcal{U}) = \frac{Q_a \times N \times M}{n}. \tag{11}$$

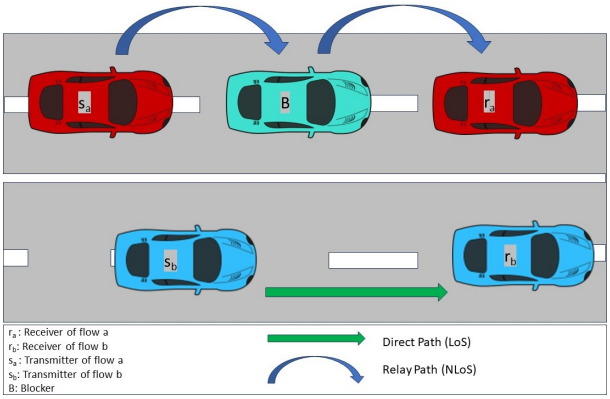

Fig. 3: Principle of Relay Selection.

V. Performance Evaluation

A. Simulation Setup

We simulate a 3 km long, 2-lane highway with 30 vehicles per lane, maintaining safety distances. Vehicles are randomly selected as transceivers, each 4.5 m long and 2 m wide. The highway width is 7.5 m, and three UAVs are positioned along the road. For performance evaluation, two metrics are considered.

- Transmission Time: The total number of time slots consumed to complete flow transmission.
- Network Throughput: The achieved throughput of completed flows in the network [Gbps].

VI. Numerical Results

The figures and tables below illustrate performance in UAV-aided environments. Fig. 4 shows time performance for PPO, DQN, and constraint programming models compared to TDMA and JRDS. Advanced algorithms, especially PPO, show significant improvements in reducing scheduling and transmission times. DQN reduced transmission time effectively but has high scheduling time. The JRDS scheme shows

TABLE I: Simulation Parameters

Parameters	Symbol	Value
Carrier Frequency	f	30 GHz
Number of flows	N	60
Number of time slots	M	2000
Slot duration	T	0.1 s
Fading depth	m	2
Background noise	N_0	-134 dBm/MHz
System bandwidth	W	2000 MHz
Transmission power	P_t	40 dBm
Average power of UAV	P_u	30 dBm
Peak power of UAV	P_u	2 P_u
Transceiver efficiency	η	0.8
Height of UAV	h_u	100 m
PL factor for V2V	α_v	2.5
PL factor of U2V	α_u	2
Rician K factor	K	9 dB
Throughput threshold	Q_a	0.1 Gbps
maximum antenna gain	G_0	21 dBi

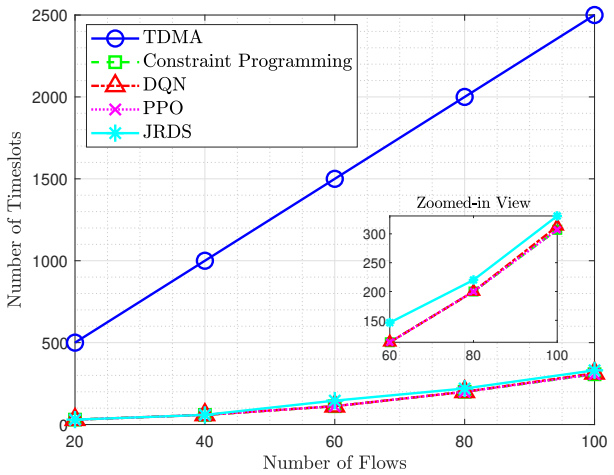

Fig. 4: Time Performance in UAV-Aided Scenario.

deviation from the ideal solution as system complexity increases.

Fig. 5 illustrates that PPO, DQN, and constraint programming outperform TDMA and JRDS in throughput. PPO and constraint programming show similar high performance, while DQN, though slightly less effective at higher flows, still surpasses TDMA. JRDS gives accurate results when scheduled flows are below 40. The resultant throughput shows great deviation from the ideal value when the number of scheduled flows are greater than 40. DQN's higher scheduling time makes it less desirable.

Fig. 6 indicates that UAV-aided scenarios complete more flows due to two-way relaying than only V2V relays. This provides enhanced flexibility and resources, with advanced algorithms leveraging these benefits.

TABLE II: Transmission Time Slots Comparison

Flows	TDMA	JRDS	Constraint Programming	DQN	PPO
20	500	30	30	30	30
40	1000	59	59	59	59
60	1500	146	112	112	112
80	2000	220	200	200	200
100	2500	331	307	313	307

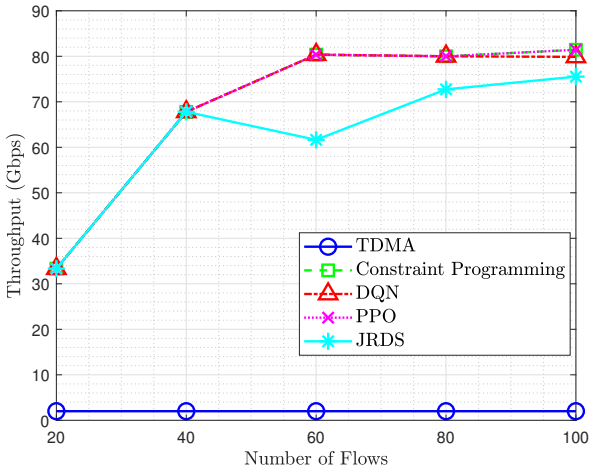

Fig. 5: Throughput in UAV-aided scenario.

A. Key Observations

- **Enhanced Scheduling:** UAVs significantly increase the capacity for scheduling data flows.
- **Optimized Throughput:** The combination of PPO and constraint programming yields optimal throughput.
- **Efficient System Performance:** PPO demonstrates the shortest scheduling and transmission times. While DQN and constraint programming offer comparable transmission times, they require longer scheduling times than PPO.

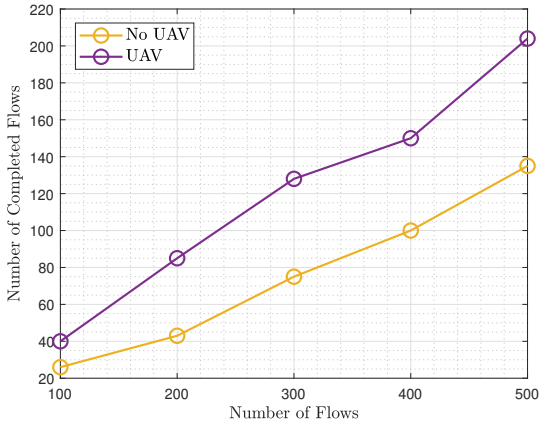

Fig. 6: Comparison of completed flows: UAV-aided vs UAV-Un-aided.

TABLE III: Network Throughput (Gbps) Comparison

Flows	TDMA	JRDS	Constraint Programming	DQN	PPO
20	2	33.33	33.33	33.33	33.33
40	2	67.79	67.79	67.79	67.79
60	2	61.64	80.36	80.36	80.36
80	2	72.73	80	80	80
100	2	75.53	81.43	79.87	81.43

VII. Conclusions

This paper shows that UAV-aided relay channels enhance V2V network performance by increasing scheduled data flows and throughput. Comparing algorithms like PPO, DQN, and constraint programming highlights their superiority over traditional methods like TDMA and JRDS. PPO, the best performer, optimizes scheduling and transmission times while adapting to dynamic environments, improving reliability and performance. Unlike rigid approaches, reinforcement learning models like PPO and DQN offer flexibility and scalability for better real-time decision-making. Future work will focus on further optimization and real-world testing.

References

[1] M. Khabbaz, J. Antoun, and C. Assi, "Modeling and performance analysis of uav-assisted vehicular networks," *IEEE Tran. Veh. Technol.*, vol. 68, no. 9, pp. 8384–8396, Sep. 2019.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[3] S. Wu, R. Atat, N. Mastronarde, and L. Liu, "Improving the coverage and spectral efficiency of millimeter-wave cellular networks using device-to-device relays," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2251–2265, 2018.

[4] C. G. Ruiz, A. Pascual-Iserte, and O. Muñoz, "Analysis of blocking in mmwave cellular systems: Characterization of the los and nlos intervals in urban scenarios," *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 16 247–16 252, 2020.

[5] J. Li, Y. Niu, H. Wu, B. Ai, R. He, N. Wang, and S. Chen, "Joint optimization of relay selection and transmission scheduling for uav-aided mmwave vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6322–6334, 2023.

[6] M. Hadded, P. Muhlethaler, A. Laouiti, R. Zagrouba, and L. A. Saidane, "TDMA-Based MAC Protocols for Vehicular Ad Hoc Networks: A Survey, Qualitative Analysis, and Open Research Issues," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2461–2492, 2015.

[7] J. Qiao, L. X. Cai, and X. Shen, "Multi-hop concurrent transmission in millimeter wave wpans with directional antenna," in *2010 IEEE International Conference on Communications*, 2010, pp. 1–5.

[8] Y. Wang, H. Wu, Y. Niu, Z. Han, B. Ai, and Z. Zhong, "Coalition game based full-duplex popular content distribution in mmwave vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 836–13 848, 2020.

[9] Q. Zhu, C.-X. Wang, B. Hua, K. Mao, S. Jiang, and M. Yao, *3GPP TR 38.901 Channel Model*. John Wiley & Sons, Ltd, 2021, pp. 1–35.

[10] Y. Zhang, L. Qiu, G. Chen, and X. Liang, "Analysis of area spectral efficiency in d2d underlaid downlink cellular networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–5.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[12] M. K. Simon and M.-S. Alouini, "Digital communications over fading channels [book review]," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3369–3370, 2008.