DDPG-based Wireless Resource Allocation for Time-Constrained Applications

Hang Hu[1], Marco Hernandez[2, 6], Yang G. Kim[3], Kazi J. Ahmed[4], Kazuya Tsukamoto[5] and Myung J. Lee[1]
[1]Department of Electrical Engineering, City College of New York, New York, NY, USA, 10031
[2]Center for Wireless Communications, University of Oulu, Oulu, Finland, 90014
[3]Department of Computer Engineering Technology, NYCCT, Brooklyn, NY, USA, 11201
[4]Department of Electrical & Computer Engineering Technology, NYIT, New York, NY, USA, 10023
[5]Department of Computer Science & Electronics, Kyushu Institute of Technology (KIT), Fukuoka, Japan, 804-8550
[6]Yokosuka Research Park-International Alliance Institute, Kanagawa, Japan, 239-0847
Emails: {hhu002@citymail, yakim@citytech, mlee@ccny}.cuny.edu, marco.hernandez@ieee.org,
kahmed17@nyit.edu, tsukamoto@csn.kyutech.ac.jp

2024 IEEE Wireless Communications and Networking Conference (WCNC) | 979-8-3503-0358-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/WCNC57260.2024.10570841

***Abstract*—This paper presents a novel model-free resource allocation framework for the downlink of 5G cellular networks to guarantee stringent QoS requirements in wireless applications. A Deep deterministic policy gradient (DDPG) agent with a modified Genetic Algorithm (GA) based resource allocation framework is proposed to balance the tradeoffs between reliability, latency, and data rate. Any feasible point in the rate-latency-reliability domain can be achieved with this approach. Compared to state-of-the-art approaches DDPG-Dual and DDPG-PSO, the proposed model achieves higher reliability and scalability in joint optimization with QoS constraints. Specifically, the proposed model guarantees the expected reliability with 25% and 42.86% improvement respectively over the compared models. In terms of conventional effective bandwidth approach, the proposed model achieves 30.82% improvement of energy efficiency under the same QoS constraints. Moreover, the proposed model offers a practical solution, namely, three times faster convergence and only 6.7% of the scheduling time compared to the ground truth Dual decomposition optimization.**

***Index Terms*—5G and beyond, Resource allocation, Deep reinforcement learning, Deep deterministic policy gradient (DDPG), Time-constrained traffic.**

I. Introduction

The emergence of time-constrained applications [1], such as industrial automation, virtual/augmented reality (VR/AR), and intelligent transportation systems (ITS) [2]–[6], has brought about stringent Quality-of-Service (QoS) requirements for 5G and beyond cellular networks. Time-constrained traffic in 5G has specific QoS requirements, including delay, reliability, data rate, etc. However, existing scheduling algorithms were not developed for time-constrained traffic but instead focused on maximizing spectrum or energy efficiency [7] by optimizing radio resources. To address this challenge, it is necessary to develop new wireless schedulers capable of meeting the stringent QoS requirements of time-constrained traffic in 5G and beyond.

Optimization algorithms and machine learning algorithms are two primary types of approaches that can be used to solve the wireless scheduling problem. Optimization algorithms face challenges when dealing with non-convex problems that lack closed-form expressions, making them suitable only for small-scale problems and delay-tolerant services. When the problem scale increases and real-time demands arise, machine learning algorithms become a very promising solution to finding optimal resource allocation in real time. This scheduling problem can be formulated as a Markov Decision Process (MDP) optimal control problem, solvable using Reinforcement Learning (RL). However, the curse of dimensionality poses a challenge for classic RL and dynamic programming when the observation/action space dimensions are high. To overcome this challenge, the actor-critic framework [8] of deep reinforcement learning has been developed. When the policy of choosing the next action is deterministic, the actor-critic deep reinforcement learning (DRL) algorithms become the deep deterministic policy gradient (DDPG) algorithm [9].

Wireless communication is susceptible to attacks that can compromise the confidentiality and integrity of transmitted data [10]. Therefore, security must be a critical consideration in the design of future wireless network systems. Physical-layer security is a promising solution to defend against eavesdropping attacks. Prior work [11] has investigated the maximum secret communication rate over an eavesdropping channel subject to reliability and secrecy constraints in the short blocklength regime. However, research on the interaction between cryptography security and wireless resource scheduling is non-existent. The 5G New Radio (NR) Packet Data Convergence Protocol (PDCP) protocol, which is responsible for security functions such as ciphering and integrity in the user's data plane, operates in layer 2 [12]. Cross-layer design [13] takes into account the delay components and packet losses in different layers, making it possible to achieve the target E2E delay and overall packet loss probability in time-constrained applications.

Achieving high reliability and low-latency communications while considering data rates in time-constrained applications poses many challenges [14]. Firstly, the communication environment is typically nonstationary, with which the pre-trained scheduling policy may fail to achieve the desired performance. Second, low latency, high reliability, and high data rate may become incompatible design parameters given the limited

wireless radio resources. Therefore, new designs are needed to balance rate-latency-reliability for time-constrained applications [15]. Third, maintaining high reliability, low latency, and high rate needs timely and efficient resource schedulers. This poses a new challenge for implementing learning-based schedulers in real-world 5G and beyond systems. Motivated by these challenges, our main contributions in this paper are summarized as follows:

- A novel model-free resource allocation framework is proposed to balance the incompatible tradeoffs between reliability, latency, and data rate without prior knowledge of users' traffic. This framework addresses the limitation of DRL when dealing with large action spaces and does not limit the problem's action space.
- To address the issue of large action spaces and enhance the scalability of DDPG, we introduce an innovative action space reduction approach, i.e., a modified heuristic Genetic Algorithm (GA) in conjunction with a Fast Water-filling algorithm (FWF) to efficiently reduce the action space in the considered wireless network. This can help accelerate convergence and improve the overall performance of the DDPG agent.
- To enhance the accuracy of latency in wireless network communication, our proposed framework incorporates radio resource optimization in a cross-layer manner for AES encryption and decryption processing. Specifically, the time overhead of user data encryption/decryption in the PDCP layer can be taken into consideration.
- Comprehensive comparisons have been conducted with state-of-the-art solutions, such as Dual decomposition, Particle Swarm Optimization (PSO), and effective bandwidth algorithms. Our proposed algorithm can stabilize the data rate given by the DDPG agent, stabilize the scalability over different users, and converge quickly.

II. System Model and Problem Formulation

A. System Model

The downlink of an Orthogonal Frequency Division Multiple Access (OFDMA) cellular network is considered, where K randomly moving users are served by one Base Station (BS) with N available resource blocks (RBs). The set of users and RBs are denoted as $\mathcal{K} = \{1, 2, .., K\}$ and $\mathcal{N} = \{1, 2, ..., N\}$, respectively. For all types of services, packets in the buffer are served according to the first-in-first-out (FIFO) order. In each time slot t, the position of each user is assumed to be fixed, and BS allocates the resource to each user according to the DDPG agent. Subsequently, each user randomly moves to their next position in the next time slot. The duration of one slot is the transmission time interval (TTI) and is denoted by Δt. It is assumed that the BS can obtain the instantaneous channel gain and the wireless environment remains unchanged during Δt.

B. Time-Constrained Traffic

The statistical QoS requirement of time-constrained traffic is characterized by the E2E delay, data rate, and overall packet loss probability [15]. The latency experienced by packets should be less or equal to the E2E delay bound to meet the delay requirement, denoted as D_k^{max} for user k. This delay comprises transmission delay, queuing delay, and processing delay. The reliability $\gamma_k(t)$ of user k at time slot t is defined as the probability of the E2E instantaneous packet delay exceeding a predefined maximum E2E latency threshold D_k^{max} [16]. Since the delay requirement of time-constrained traffic in this paper is much longer than the channel coherence time, the decoding errors can be ignored [17]. To circumvent extended transmission delays, reliability cannot be improved via retransmission. The achievable rate of the k-th user can be expressed as:

$$R_k(t) = \sum_{n=1}^{N} \rho_{k,n}(t) W \log_2 \left(1 + \frac{\alpha_k(t) g_{k,n}(t) p_{k,n}(t)}{\sigma^2}\right), \tag{1}$$

where W represents the bandwidth of each RB. $\alpha_k(t)$ is the large-scale channel gain of user k at slot t, and $g_{k,n}(t)$ is the small-scale channel gain of the transmission from the BS to user k on RB n at time slot t. We use indicators $\rho_{k,n}(t)$ to represent whether RB n is scheduled to user k at time slot t. If RB n is allocated to user k at time slot t, $\rho_{k,n}(t) = 1$. Otherwise, $\rho_{k,n}(t) = 0$. Based on OFDM principles, each RB can only be occupied by one user at time slot t. $p_{k,n}(t)$ is the downlink transmission power of the BS to user k on RB n at time slot t. σ^2 is the noise power and it equals WN_0, where N_0 is noise spectral density.

Our research includes the processing delay of user data encryption/decryption of the PDCP layer in the formulation. Therefore, the time overhead of cryptographic processing can be reflected in E2E latency. Our proposed optimization framework can optimize radio resources in a cross-layer manner, allowing us to consider the delay components across different layers. Hence, it is possible to achieve the E2E target delay in time-constrained traffic.

C. Problem Formulation

This paper aims to allocate resources that minimize the transmission power while ensuring users achieve their target data rate, reliability, and cross-layer latency. The optimization problem is formulated as below. The objective function in (2a) is formulated to optimize transmission power through the RB assignment indicator $\rho_{k,n}(t)$ and the allocated power $p_{k,n}(t)$. In (2b), D_k is the packet delay of user k, and γ_k^* is the target reliability of user k. Constraint (2b) explicitly accounts for the latency and reliability of each user, guaranteeing that the E2E latency is less or equal to D_k^{max} with minimum reliability of γ_k^*. A model-free DRL agent is proposed that can adaptively learn from practical users' traffic without prior knowledge of the traffic model. Specifically, we calculate the reliability in (2b) using empirical measurements [15], i.e., $\gamma_k(t) = \Pr\{D_k \leq D_k^{max}\} \approx 1 - \frac{\nu_k'(t),}{\nu_k(t)}$, where ν_k' is the number of packet transmission to user k in time slot t, whose E2E delay exceeds D_k^{max}, while $\nu_k(t)$ is the total number

of packets transmission to user k in time slot t. A cross-layer framework is established for the downlink transmission of time-constrained service, where E2E delay comprises the transmission delay in the PHY layer, the queuing delay and the scheduling delay in the MAC layer, and the encryption/decryption delay in the PDCP layer. Moreover, users can convey their E2E delay and the received number of packets to the BS via the uplink control channel. Constraint (2c) explicitly captures the data rate constraint. $\mu(\cdot|\theta^{\mu})$ represents the actor of DDPG, and θ^{μ} represents the parameters of the actor, i.e., weights and biases. $\beta_k(t)$ is the average transmitted packet size for each user k. Deep Neural Network (DNN) universal function approximators [18] are used to estimate the action given a system state. Constraint (2d) is the minimum data rate requirement of user k, which can vary for different applications. Constraint (2e) is the feasibility of the solution.

$$\mathbf{P1}: \quad \underset{p_{k,n}(t),\rho_{k,n}(t)}{\text{minimize}} \quad \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}(t), \tag{2a}$$

$$\text{s.t.} \quad \Pr\{D_k \le D_k^{max}\} \ge \gamma_k^*, \tag{2b}$$

$$R_k(t) \ge \mu(\nu_k(t), \beta_k(t)|\theta^{\mu}), \tag{2c}$$

$$R_k(t) \ge R_{min}, \tag{2d}$$

$$p_{k,n}(t) \ge 0, \quad \rho_{k,n}(t) \in \{0,1\}, \quad \sum_{k=1}^{K} \rho_{k,n}(t) = 1,$$
$$\forall k \in \mathcal{K}, \quad \forall n \in \mathcal{N}, \quad \forall t. \tag{2e}$$

In summary, the resource allocation scheduler has two functions. The first function is to estimate the data rate of each user to guarantee that the reliability and latency requirements are met. According to the estimated data rate, the second function is to select the assignment RB indicator and allocates power to each user, minimizing the total transmission power. A noble action space reduction is proposed to facilitate this process, which will be explained later.

III. Resource Allocation with DDPG

A. DDPG for Scheduler Design

This section proposes a DDPG framework to solve the problem **P1** as discussed in section II. DDPG is a good choice since it is continuous rate control and more sample-efficient based on deterministic policy. An RL problem typically consists of three components: state space $\mathcal{S}$, action space $\mathcal{A}$, and reward $\mathcal{R}$. At each state $s_t \in \mathcal{S}$, our DDPG agent takes action $a_t \in \mathcal{A}$, receives an immediate reward r_t and moves to the next state s_{t+1}. These sequences are referred to as one transition, and it is stored in a replay memory with the size of $|\mathcal{D}|$ as the training data. For our wireless resource allocation problem, these components are defined as follows.

States: At each time slot t, the state comprises the channel gain $c_{k,n}(t)$, the number of packets $\nu_k(t)$ transmitted to each user, and the average packet length $\beta_k(t)$ for each user k. This channel gain incorporates large-scale path loss $\alpha_k(t)$ and small-scale fading $g_{k,n}(t)$. The state is defined as: $s_t = \{c_{k,n}(t), \nu_k(t), \beta_k(t)\}$.

User mobility is modeled as random movement around BS, which results in the distance variation between users and BS. These distance variations are then mapped into channel gain variations. It only affects the states and does not impact the design of our framework. To address the challenge of the large state space problem due to continuous states, we employ deep Q-network. Furthermore, a novel heuristic mechanism, a modified GA, is proposed to reduce the action space.

Actions: The action space $\mathcal{A}$ contains all the possible decisions for assigning RBs $\rho_{k,n}(t)$ and the power allocation $p_{n,k}(t)$. Thus, the action of the scheduler at time slot t is defined as: $a_t = \{\rho_{k,n}(t), p_{n,k}(t)\}$.

Rewards: The immediate reward is represented using the overall transmit power and the measured reliability of each user. The reward is defined as follows: $r_t = \sum_{k\in\mathcal{K}, n\in\mathcal{N}} [-w_k(t)(1-\gamma_k(t)) - c_1 p_{k,n}(t)]$, where c_1 is a weighting factor for power and $w_k(t)$ is a time-varying weight that ensures reliability over time slots as the network states change dynamically. $w_k(t)$ is given in detail by: $w_k(t+1) = \max\{w_k(t) + \gamma_k^* - \gamma_k(t), 0\}$. Failure to meet the required reliability results in an increase of the corresponding time-varying weight factor. Hence, the weighting factor ensures that the system meets the target reliability of the users.

Fig. 1 illustrates the structure of the DDPG framework designed for the downlink OFDM system. The DDPG network based on the actor-critic framework consists of four networks: the actor network $\mu(s|\theta^{\mu})$ responsible for selecting actions, the critic network $Q(s,a|\theta^{Q})$ for estimating Q values of selected actions, the corresponding target actor network $\mu'(s|\theta^{\mu'})$ and target critic network $Q'(s,a|\theta^{Q'})$ utilized for generating the target values for training, where $\theta^{\mu'}$, θ^{Q} and $\theta^{Q'}$ represent the parameters of the networks. The DDPG uses the Bellman equation to optimize the parameters of the critic. Consequently, the critic network updates its weights θ^{Q} by minimizing the loss function: $L(\theta^{Q}) = \frac{1}{N_{tr}} \sum_{i=1}^{N} (y_i - Q(s_i, a_i|\theta^{Q}))^2$, where $y_i \triangleq r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ and N_{tr} denotes the batch size. To address the unstable learning issue caused by using only a single network, the target networks are updated softly. Specifically, the actor network updates its weights θ^{μ} in the direction of policy gradient, which can be expressed as: $\nabla_{\theta^{\mu}} J \approx \frac{1}{N_{tr}} \sum_{i=1}^{N} \left(\nabla_{\theta^{\mu}} \mu(s|\theta^{\mu})|_{s_i} \nabla_a Q(s,a|\theta^{Q})|_{s=s_i, a=\mu(s_i)} \right)$.

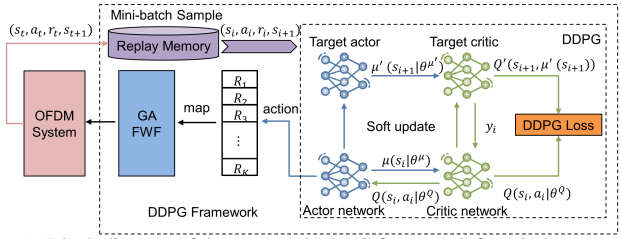

Fig. 1. Block diagram of the proposed DDPG framework for wireless resource allocation

B. Action Space Reduction

OFDM allows one user to occupy multiple RBs, and the transmit power applied on each RB is continuous, resulting in $\mathcal{O}(K^N) \times \mathbb{R}^N$ mixed integer action space size. Consequently,

the optimization problem is classified as Mixed Integer Non-Linear Programming (MINLP), which is known to be NP-complete. Moreover, employing this large action space directly in DRL algorithms is challenging. To solve this problem, we use the DDPG algorithm to estimate the data rate for each user, which reduces the original action space to a small continuous action space $\mathbb{R}^K$. Subsequently, we use a modified heuristic Genetic Algorithm (GA) that incorporates a Fast Water-filling Algorithm (FWF) to map the actions taken by the DDPG algorithm to the original action space. In the absence of GA-FWF mapping, the DDPG agent is unable to translate its outputs into viable practical actions. The modified GA is applied to solve the RB assignment problem, while the FWF algorithm is used to address the power minimization problem. The merit of the modified GA and FWF combination can guarantee high efficiency and performance as shown in the performance analysis part.

1) Modified Genetic Algorithm for RB Assignment: Inspired by the natural selection process, the Genetic algorithm [19] can effectively solve the mixed integer optimization problem and is suitable for integer RB assignment optimization in OFDMA. Specifically, the GA generates chromosomes with N elements and sets the total population P. The value of each element in the chromosome is confined to the integer from 0 to $K-1$, which represents the users. For example, the value of n-th element in the chromosome is k, meaning that the n-th RB is allocated to user k. The element of the chromosome is randomly generated, and each chromosome represents a potential solution to the optimization problem.

The water-filling algorithm is employed to calculate the overall transmit power as the fitness function. The goal is to minimize the transmission power, and therefore, lower power consumption leads to higher fitness of the chromosome. Elitist selection and single-point crossover are applied to determine the offspring for the subsequent generations. The mutation operation randomly changes the bits in a chromosome following Gaussian distribution.

More importantly, a modified GA is applied to improve the performance of RB assignment by introducing an individual with *good* genes to the initial population. This approach considers two aspects: efficiency and fairness. Specifically, the number of RBs required by each user is determined according to their rate requirement R_k, and the RBs are allocated to the user with the largest channel gain at the same RB. The computational complexity of the GA is $\mathcal{O}(GP^2)$, where G denotes the generation, and P is the total population. This is a significant reduction in computation compared with $\mathcal{O}(K^N)$ complexity.

2) Fast Water-filling Algorithm for Power Minimization: The modified GA can provide a feasible RB assignment and then the Fast Water-filling algorithm (FWF) [20] is utilized to allocate transmission power to each user based on their assigned RBs. The process continues until the stopping criteria of the modified GA and the optimal RB and power allocation vectors will be obtained. It is worth noting that the water-filling algorithm minimizes transmit power while ensuring a target rate for a single user, given the RB assignment. It was shown in [20] that the computational complexity of the conventional water-filling algorithm is N^2+2N+5. However, the Fast Water-filling algorithm reduces the complexity to $3N+4$.

IV. Performance Analysis

Our simulation platform is implemented in the Python framework using PyTorch and Tianshou [21], designed to facilitate the implementation of deep reinforcement learning models. We trained and evaluated our DDPG agent in a customized environment based on Nokia Bell Lab's open-source Wireless Suites simulator [22].

A. Simulation Platform

An OFDMA cellular system with 50 RBs serves 15 users in our simulation platform. Users move at random speeds along random rectilinear trajectories within a 1 km x 1 km square area. At the start of each episode, we randomly position each user within the area. The channel model includes large-scale path loss α_k and small-scale fading $g_{k,n}$. The path loss exponent is set to 3 (urban area), and small-scale fading follows the Rayleigh fading. Additionally, the packet arrival process of each user follows Poisson process. The inter-arrival time between packets follows exponential distributions with parameters λ. Moreover, the DRL agent can adaptively learn from practical users' traffic without prior knowledge of the traffic model. The simulation setup is summarized in Table I.

B. Latency Analysis of Security Processing

The processing delay for security is measured experimentally for time-constrained services. AES encryption with GCM mode is implemented based on the pyca/cryptography library [23]. The result shows that the encryption time for 10 kbits of data with a 256-bit key is approximately 0.116 ms. In our downlink scheduler, user data traffic is encrypted at BS before transmission over the air interface. Typically, BS has powerful servers that can support sufficient computation quickly. However, the decryption of user data is performed on the UE side with mobiles or IoT devices, which have limited computation resources. Based on this fact, we estimate that the decryption time on the UE side is 20 times longer than our simulation results [24], and the latency is about 2.32 ms. Since the TTI Δt is set to 1 ms in this paper, the security latency is approximated to be 3 ms.

C. Experiments for Action Space Reduction

In the DDPG framework (Fig. 1), the modified GA employs the FWF algorithm as its fitness function for total transmission power, with lower fitness values indicating preference. The implementation of the GA is based on the open-source scikit-opt library [25], a Python module for heuristic algorithms. To ensure fairness in data rate and efficiency of channel gain, we include an individual with *good* genes in the initial population. Meanwhile, a significant penalty is added to the chromosome if it fails to meet the fairness criteria for data rate. The chromosome size is set at 100, and the mutation probability is set to 0.001.

TABLE I
SIMULATION PARAMETERS

System setup	
Carrier frequency	2 GHz
Total bandwidth B	10 MHz
Bandwidth of each RB W	180 kHz
Maximum BS power P_{max}	10 W
Noise power spectrum density σ^2	-173.9 dBm/Hz
Packet size $1/\nu^s$	10 kbits
Time slot interval Δt	1 ms
Maximum latency D_i^{max}	13 ms
Security latency	3 ms
Packet arrival rate λ	0.1 packet/ms
Learning setup	
Replay memory size $\|\mathcal{D}\|$	10000
Actor learning rate	10^{-5}
Critic learning rate	10^{-4}
Discount factor	0.9
Soft update rate τ	10^{-3}
Exploration noise	1
Batch size N_{tr}	64
Time slots per episode	100
Hidden layer size	[512, 256, 128, 64, 32]

We compare our action space reduction with two other baseline approaches, namely, Particle Swarm Optimization (PSO) algorithm [26] and iterative Dual decomposition algorithm [15]. The Dual decomposition algorithm represents a closed-form solution specifically crafted to attain optimal power allocation, yet it is afflicted by scalability issues. For simplicity, the data rate is fixed at 1 Mbps. We record the minimum fitness during the iterations, and the iterative convergence of three algorithms is depicted in Fig. 2. It is evident that our modified GA converges after only 18 iterations, whereas the iterative Dual decomposition algorithm requires several orders of magnitude more iterations to converge.

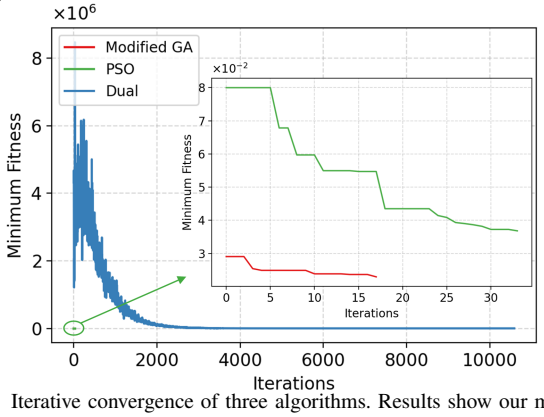

Fig. 2. Iterative convergence of three algorithms. Results show our modified GA converges faster than the other two baselines

Fig. 3(a) shows the processing time of three algorithms with six different stopping criteria, i.e., 10^{-1} to 10^{-6}. For the parameters of PSO, we set the weight factor for inertia to 0.8, and self-confidence and swarm-confidence to 1.5 each. The ellipsoid method is applied in the iterative Dual decomposition, with the initial ellipsoid set as $(2 \times 10^3$, rate error), where the rate error is the difference between the actual rate and the desired rate. The experiments were repeated six times, and the 95% confidence interval was calculated using the t-distribution. The result demonstrates that our modified GA achieves the shortest processing time at each stopping criterion compared to the other two baselines. Fig. 3(b) shows the total transmission power of the three algorithms, where the Dual optimization achieves the optimal transmission power, and our modified GA achieves solutions very close to the optimal one.

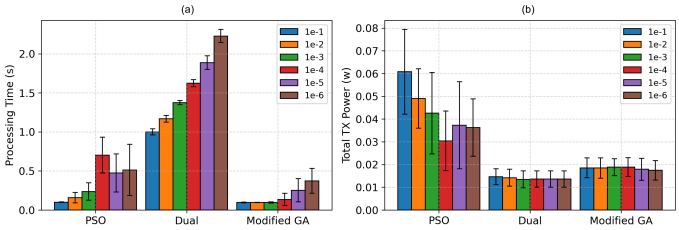

Fig. 3. (a) Processing time of three algorithms. (b) Total transmit power of three algorithms

According to the results, we conclude that PSO is unsuitable for such integer problems as RB assignment, as it does not perform well in power optimization. Importantly, while the Dual decomposition algorithm can achieve optimal power consumption, its actual data rate fluctuates and cannot always guarantee the desired rate for each user, as shown in Fig. 4 (a). In contrast, our modified GA can consistently guarantee the desired rate. On the other hand, our modified GA and PSO can keep the processing time at a low level. However, Dual decomposition has a severe scalability problem, which means that when the number of users increases, the processing time will increase significantly, as shown in Fig. 4 (b). In summary, our modified GA can balance the advantages of two aspects, i.e., time consumption and power minimization, while guaranteeing the desired rate.

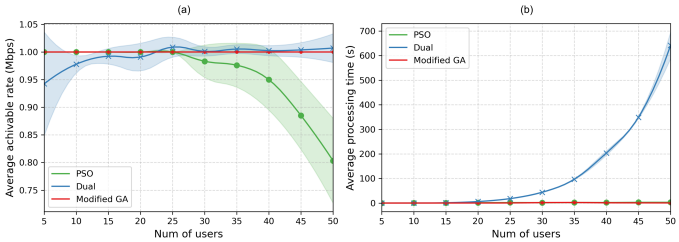

Fig. 4. (a) The achievable data rate for different numbers of users (b) The average processing time for different numbers of users

D. System Performance

Having analyzed the performances of individual components of the system (Fig. 1), the system level performance is evaluated in this section. The desired reliability is set to 99%, and the action space range is defined as [1, 10] Mbps, with a minimum data rate requirement of 1 Mbps. Considering the computation time and the consistent performance order of the algorithms over the range of stopping criteria shown in Fig. 3, the stopping criterion is set to 10^{-3}. Before assessing the learned policy's performance, we first evaluate the convergence of the proposed algorithm and the benchmark algorithms under the same parameter configurations, as shown in Table I. Fig. 5 illustrates the average testing rewards of four methods. It is evident from Fig. 5 that both DDPG-GA and the baseline algorithms gain more experience as training progresses and collect higher expected average rewards. DDPG-Dual achieved better testing rewards than those of DDPG-GA and DDPG-PSO as expected. However, the DDPG-Dual, which involves 200 epochs, took about 15 hours to train, whereas our proposed DDPG-GA took only about 5 hours, one-third of the training time of DDPG-Dual.

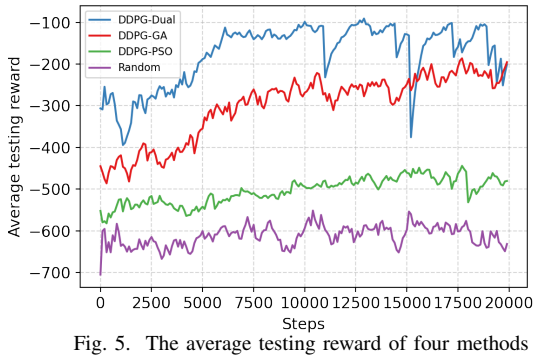

Fig. 5. The average testing reward of four methods

To assess the reliability performance and energy efficiency of the learned policy, six repeated measurements were conducted. Table II shows all users' minimum and maximum reliability and average energy efficiency in the measurements. The effective bandwidth is a traditional method to express the minimum constant service rate required by a given time-varying arrival process to guarantee a probabilistic delay constraint. The size of packet follows exponential distribution with parameters ν^s. Then, the effective bandwidth (EB) of the k-th user can be expressed as follows [27] $E_k^B = \frac{\lambda}{\nu^s - q_k}$, (bits/s), where q_k is the QoS exponent, which can be obtained from $e^{-q_k E_k^B(q_k) D_k} \approx (1 - \gamma_k^*)$. The results indicate that the proposed DDPG-GA and traditional Effective Bandwidth GA (EB-GA) achieved a reliability of 100%, while the minimum reliability of DDPG-Dual and DDPG-PSO were 75% and 57.14%, respectively, which failed to guarantee the desired reliability of 99%. In terms of average energy efficiency, our proposed DDPG-GA achieved better results than traditional EB-GA approach.

TABLE II
RELIABILITY ANALYSIS (RA) AND ENERGY EFFICIENCY

Algorithms	DDPG-Dual	DDPG-PSO	EB-GA	**DDPG-GA**
RA (%, min/max)	75/100	57.14/100	100/100	**100/100**
EE (Mbps/W)	3.75	0.83	1.59	**2.08**

V. CONCLUSION

In summary, this paper provides a more practical solution to address the wireless resource allocation problem for time-constrained traffic in the downlink OFDMA system. A DDPG-based scheduler is designed to allocate RBs and transmission power while meeting users' required reliability, latency, and data rate. To solve the issue of large action space, we proposed a modified GA that can accelerate convergence. In addition, our proposed model can achieve stable data rates given by the DDPG agent, which is more helpful in improving reliability. The system performance indicates that our proposed model can guarantee the reliability with 25% and 42.86% improvement and demonstrates faster convergence and scheduling time compared to the other state-of-the-art algorithms. Under the same QoS constraints, average energy efficiency of our proposed model achieves 30.82% improvement compared to the traditional effective bandwidth method.

On the other hand, the potential limitation of the proposed algorithm is that when the environment is highly dynamic (e.g., mobility pattern, channel condition), the predetermined hyper-parameters of the actor, the critic, and modified GA might deteriorate the QoS performance. As future work, we intend to investigate alternative DRL approaches, including stochastic policy-based algorithms, and conduct comparative analyses with deterministic policies. Subsequently, our objective is to implement the proposed methodology on the NSF COSMOS wireless testbed for practical deployment.

ACKNOWLEDGMENT

This work is supported by NSF PAWR COSMOS (#1827923) and NSF IRNC COSMIC (#2029295) grants.

REFERENCES

[1] Study on Scenarios and Requirements for Next Generation Access Technologies, document 3GPP TSG RAN TR38.913 R16, Jul. 2020.
[2] P. Schulz et al., "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," IEEE Commun. Mag., 55(2), 70-78, 2017.
[3] C. She et al., "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," Proc. IEEE, 109(3), 204-46, 2021.
[4] H. Hu, M. J. Lee, "Graph Neural Network-based Clustering Enhancement in VANET for Cooperative Driving," ICAIIC, pp. 162-167, 2022.
[5] T. J. Chen et al, "Open-access millimeter-wave software-defined radios in the PAWR COSMOS testbed: Design, deployment, and experimentation," Computer Networks, 2023, Jul, 13:109922.
[6] A. Qadeer, M. J. Lee, "HRL-Edge-Cloud: Multi-Resource Allocation in Edge-Cloud based Smart-StreetScape System using Heuristic Reinforcement Learning," Inf Syst Front, 1-17, 2023.
[7] M. Amjad, L. Musavian, M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," IEEE Commun. Surveys Tuts., 21(4), 3007-38, 2019.
[8] D. Silver, et al., "Deterministic policy gradient algorithms," International conference on machine learning, pp. 387-395, Pmlr, 2014.
[9] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," arXiv:1509.02971, 2015.
[10] C. Li et al., "5G-based systems design for tactile Internet," Proc. IEEE, 107(2), 307-24, 2018.
[11] W. Yang, R. F. Schaefer, H. V. Poor, "Wiretap channels: Nonasymptotic fundamental limits," IEEE Trans. Inf. Theory, 65(7), 4069-93, 2019.
[12] Security architecture and procedures for 5G System, document 3GPP TS 33.501 V17.6.0 R17, 2022.
[13] C. She et al., "Cross-layer design for mission-critical IoT in mobile edge computing systems," IEEE Internet Things J., 6(6), 9360-74, 2019.
[14] M. Bennis, M. Debbah, H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," Proc. IEEE, 106(10), 1834-53, 2018.
[15] A. T. Kasgari et al., "Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication," IEEE Trans. Commun., 69(2):884-99, 2020.
[16] Service requirements for the 5G system, 3GPP, TS 22.261, R16, 2021.
[17] R. Dong et al., "Deep learning for radio resource allocation with diverse quality-of-service requirements in 5G," IEEE Trans. Wireless Commun., 20(4), 2309-2324, 2020.
[18] Y. Lu, J. Lu, "A universal approximation theorem of deep neural networks for expressing probability distributions," Advances in neural information processing systems, 33:3094-105, 2020.
[19] Y. Wang, F. Chen, G. Wei, "Adaptive subcarrier and bit allocation for multiuser OFDM system based on genetic algorithm," In Proc. ICCCS, Vol. 1, pp. 242-246, 2005.
[20] S. K. Taskou, M. Rasti, "Fast water-filling method for sum-power minimization in OFDMA networks," IEEE Trans. Signal Process., 24(7), 1058-1062, 2017.
[21] J. Y. Weng et al., "Tianshou: A highly modularized deep reinforcement learning library," arXiv:2107.14171, 2021.
[22] "Wireless-suite," Nokia, Available: https://github.com/nokia/wireless-suite, 2021.
[23] pyca/cryptography. https://github.com/pyca/cryptography. Latest release: 3.4.8. Last accessed: August 24, 2021
[24] X. Wang et al., "Performance evaluation of attribute-based encryption: Toward data privacy in the IoT," ICC, pp. 725-730, 2014.
[25] guofei9987. scikit-opt. https://github.com/guofei9987/scikit-opt, 2020.
[26] Y. Z. Zhang et al., "PSO-based minimum residual algorithm for mobile robot localization in indoor environment," IJARS, 14(5), 2017.
[27] F. P. Kelly, "Notes on effective bandwidths," Stochastic Networks: Theory and Applications. London, U.K.: Oxford Univ. Press, 1996.