Dynamic Subcarrier Assignment in OFDMA-PONs Based on Deep Reinforcement Learning

Min Zhu, *Member, IEEE*, Jiahua Gu, Bin Chen, and Pingping Gu

***Abstract*—Orthogonal Frequency Division Multiplexing Access Passive Optical Networks (OFDMA-PONs), a solution for the next-generation optical access network, allows multiple optical network units (ONUs) to dynamically share subcarriers (SCs) to support efficient bandwidth allocation. In uplink transmission, multiple ONUs can share orthogonal low bit rate SCs to transmit data at different time slots (TSs) during the transmission cycle. In this paper, the dynamic subcarrier allocation (DSA) scheme based on deep reinforcement learning (DRL) is proposed for various ONU bandwidth requests. The novel scheme jointly allocates time slots, subcarriers, and modulation formats in a dynamic and flexible manner. The ONU can save transmit power by using a lower order modulation format while meeting the delay requirement. The simulation part demonstrates how the proposed DRL-based DSA scheme can be adapted to various situations, including 1) variation in the size of ONU bandwidth requests, and 2) variation in the weight of different indicators. The extensive simulation results show that, for the first time, the proposed DRL-based DSA algorithm achieves optimal traffic latency with substantial power saving, compared with the traditional two-dimensional DSA algorithms.**

***Index Terms*—Energy-saving, deep reinforcement learning, dynamic subcarrier assignment (DBA), orthogonal frequency division multiplexing access passive optical network (OFDMA-PON).**

I. Introduction

With the ever-increasing demand for bandwidth from various types of multimedia services, e.g., high definition video streaming, edge computing, cost- effective passive optical networks (PONs) have become mature technologies for broadband access and have been widely deployed worldwide [1]. The so-called "passive" is that there is no active element in the entire optical distribution network (ODN) of the PON from the center office (CO) to user-side premise. Advantages of using PON include large coverage area, reduced fiber deployment, multicast and broadcast capabilities, reduced cost of maintenance, and ease of upgrades to higher bit rate [2]. However, with the rapid development of network technology and digital services, to better support the high quality of service (QoS) requirements of access networks, PON needs a higher data access rate [3]–[5]. The wavelength division multiplexing passive optical network (WDM-PON) has been proposed as a candidate solution, which provides a logic end-to-end wavelength connection for each user [1]. However, the available wavelength capacity is not used effectively under the current application scenario. The optical code division multiple access PON (OCDMA-PON) designates each user a unique optical code sequence. The coded bits are transmitted over the ODN and then decoded using the exact optical code sequence at the receiver of the destined user. With the capability of multi-user wavelength sharing, the OCDMA-PON offers a more flexible bandwidth allocation and supports a larger number of users than WDM-PON [6], [7]. However, the spectrum efficiency of OCDMA-PON is lower because it is based on the spread-spectrum technique. Hence, the orthogonal frequency division multiple access passive optical network (OFDMA-PON) that enables the sharing of sub-wavelength resources in the frequency domain is proposed to address the effective bandwidth allocation [1], [3], [5], [8]–[11].

Orthogonal Frequency Division Multiplexing (OFDM) technology has recently gained remarkable development in the field of optical networks [1]. Due to the advantages of large capacity, high spectrum efficiency (SE), transparency to modulation formats, flexible multiple address access, dynamic bandwidth allocation (DBA), the OFDMA-PON has been considered as a good candidate to further support the system capacity increase for next-generation optical access network 2 (NG-PON2) [12]. Recent reports suggest that OFDMA-PON is a promising candidate for the systems beyond NG-PON2 [13], [14]. In addition, OFDMA-PON allows different optical network units (ONUs) to transmit upstream data through a set of shared orthogonal low bit rate subcarriers (SCs) in different time slots (TSs) during the transmission polling cycle to meet bandwidth requirements and realize various data rates [10], [15].

The typical situation in OFDMA-PONs is that the bit rate of each subcarrier is far lower than the average ONU data rate [1]. OFDM divides the transmitted signal into a number of low-rate SC signals, which partially overlap in the frequency domain but do not interfere (using orthogonality) [1], [11]. This means that in OFDMA-PONs, several SCs must be grouped together to provide the required bandwidth for ONUs. Dynamic subcarrier allocation (DSA) is designed to manage the allocation

Manuscript received January 14, 2022; revised January 28, 2022; accepted January 30, 2022. Date of publication February 4, 2022; date of current version February 28, 2022. This work was supported in part by Jiangsu Provincial Key Research and Development Program under Grant BE2020012 and in part by the Transformation Program of Scientific and Technological Achievements of Jiangsu Province under Grant BA2019026. *(Corresponding author: Min Zhu.)*

Min Zhu and Jiahua Gu are with National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China and also with Purple Mountain Laboratories, Nanjing 211111, China (e-mail: minzhu@seu.edu.cn; gujiahua@seu.edu.cn).

Bin Chen is with the School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: cb@seu.edu.cn).

Pingping Gu is with Taicang T&W Electronics Company Ltd., Taicang 215400, China (e-mail: gupingping@twsz.com).

Digital Object Identifier 10.1109/JPHOT.2022.3148259

of SCs for each ONU request in OFDMA-PON. Whether an ONU bandwidth request being satisfied would depend on the following two factors: 1) the modulation format adopted by the SCs, which determines the bit number transmitted per OFDM symbol; for example, when the modulation format of M-QAM is used, each OFDM symbol carries $\log_2 M$ bit [1]; and 2) the baud rate of the OFDM symbol, which is closely relevant to the available spectrum resource, i.e., how many SCs are allocated to each ONU request. In the above first point, the modulation formats allocated for an ONU request determine not only the spectrum utilization of these SCs, but also the transmitting power and traffic delay of the ONU. For instance, if the SCs allocated for an ONU request adopt the lower-order modulation format, then the ONU transmission power is lower, but the traffic delay of the ONU request would be higher, because that more SC resources required were less likely to be satisfied. By contrast, if the higher-order modulation format is used for the SC, the ONU transmission power is higher, but less SCs are needed for the ONU request due to the higher spectrum utilization.

In practical applications, minimizing the power consumption of ONU is always a priority, as it accounts for 60% to 70% of OFDMA-PON's energy consumption [15]. To realize energy-efficient DSA in OFDMA-PONs, it is essential to jointly allocate time slot, SC, and modulation format, because both the ONU transmission power and the ONU traffic delay are closely correlated. A well-designed DSA scheme needs to meet the following challenges: 1) seek the appropriate scheduling order for multiple ONU requests to reduce the traffic latency; 2) find the optimal modulation format for each SC group to reduce the transmission power.

Although the DSA problem has been investigated widely in the wireless systems, there exist much different from our works in terms of the network topology, the allocated resource and the optimization objective. For example, the allocated resource in [16] is only the radio resource blocks (RBs) to improve the cell throughput and guarantee access proportional fairness. But in our work, the allocated resources in OFDM-PON include the subcarriers, the time slots and the modulation formats and our goal aims to minimize the power consumption and the delay of ONU requests, simultaneously.

Many solutions have been proposed to reduce the cost of OFDMA-PON [1], [3], [5], [10], [11], [17]–[21]. In media access control (MAC) layer, a DSA algorithm was proposed in [1] to schedule the ONU upstream transmission based on its instantaneous bandwidth requirements and the existing traffic conditions. In [17], the energy efficiency of WDM-OFDM-PON is improved by sharing the OFDM modulation module on the physical layer and the MAC layer. Early works related to algorithm-level cost reduction mainly focuses on two-dimensional resource scheduling, i.e., time slot and SC allocation. The DSA issue of OFDMA-PON was also studied in [10], using the offline scheduling framework to analyze the SC utilization and the total granted time. In OFDMA-PON, a weighted DSA scheduling algorithm was proposed to reduce the terminal wireless data packet delay [11]. In [18], a randomized dynamic bandwidth allocation algorithm for upstream access in OFDMA-PON was proposed to improve throughput and reduce package delay. In [19], the authors proposed and experimentally demonstrated an all-optical virtual private network (VPN) supporting dynamic bandwidth allocation (DBA) in OFDMA-PON system. Authors in [20] proposed an interleaved polling with adaptive cycle time (IPACT)-based 2D bandwidth allocation method for the OFDMA-PON to guarantee delay performances for time sensitive services. A fair-aware DSA algorithm in a distance adaptive OFDMA- PON was proposed in [21]. In [5], a novel DSA algorithm is proposed that combines traffic prediction technology to reduce latency. A DSA framework based on weight distribution in heterogeneous OFDMA-PON was proposed in [22]. However, in all existing related works, neither the flexible configuration of the SC modulation format nor the ONU transmission power optimization was considered.

Recently, some solutions have been proposed to address the ONU transmission power optimization issue. In [3], the ONU transmission power is minimized by optimally allocating SC and modulation format in each TS. Joint allocation of virtual subcarriers (VS), TS, and modulation formats was studied to maximize energy savings with multi-dimensional resource re-allocation and flexible ONU re-configuration [15]. The authors in [23] proposed a distance-adaptive bandwidth allocation scheme to realize low-cost high-capacity long-range OFDMA-PONs. In [24], a number of sub-bands are grouped together as a band group (BP) and multiple ONUs share the BP by time division mode to realize an energy-efficient time division multiple band allocation passive optical network (TDMBA-PON). However, none of the above works considered the required quality of service (QoS) such as traffic delay requirement.

Recently, Deep Reinforcement Learning (DRL) has been successfully applied to some complex decision-making problems in resource management. In particular, complex systems and decision strategies can be modeled as deep neural networks (DNN), trained to achieve optimal mapping from the input (i.e., state space) to the output (i.e., action space). The application of the DRL-based algorithms in improving the performance of communication networks has recently attracted much attention from both academia and industry. Most of these efforts focus on resource scheduling problems in systems and networking. In [25], iterative point-wise reinforcement learning for highly accurate indoor visible light positioning (VLP) was proposed to reduce positioning errors. [26] studied the DRL-based slice admission policy to maximize the profits of infrastructure providers (InP). [27] addressed the multi- resource cluster scheduling problem with DRL strategy to minimize average job slowdown. A joint BBU placement and routing strategy based on the DRL in C-RAN was proposed in [28] to maximize resource utilization. In [29], a DRL-based strategy was proposed to improve the overall network performance in the elastic optical network (EON).

In this paper, we propose a DRL-based DSA algorithm that flexibly assigns SCs based on the ONU requests and the total available SCs. It jointly allocates SCs, TSs, and modulation formats to minimize the power consumption and the delay of ONU requests, simultaneously. Specifically, it determines which ONU request should be served first and which modulation format is used for the SCs assigned to each ONU request. Extensive

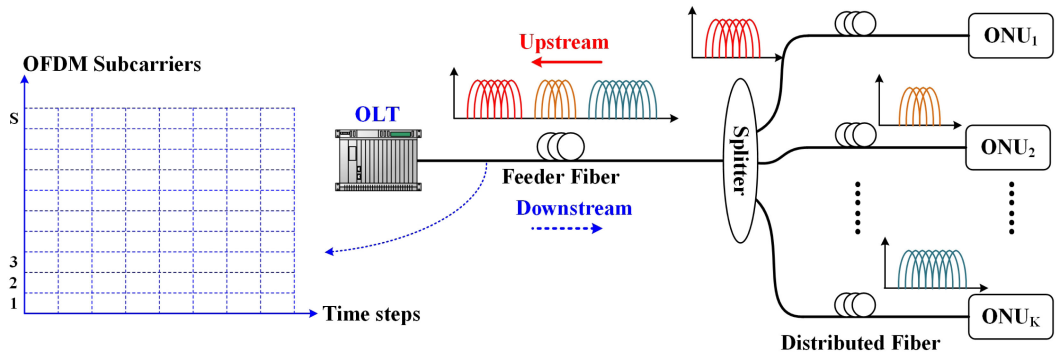

Fig. 1. OFDMA-PON architecture with S SCs and K ONUs.

numerical simulations are conducted to evaluate the performance of our proposed DRL-based DSA scheme. Note that the bassline heuristics used in the simulation part, such as Tetris, SJF, Packer and Random [30]–[32], generally use fixed policy to schedule the serving order of ONU requests. When network state changes, these fixed policies cannot adapt to the changes of the network state. Moreover, these bassline heuristics are not designed based on the required optimization goal. However, by iterative training, our proposed DRL-based scheme has the flexibility to accommodate different network states and can achieve the optimization in accordance with the specified goals, i.e., minimize the power consumption and the delay simultaneously. Simulation results show that the DRL-based DSA scheme is able to maintain low traffic latency while saving power using low modulation format and has excellent robustness against the variations in ONU demand.

The rest of the paper is organized as follows. Section II presents the system architecture and formulates the problem. Section III elaborates on the design of the DRL model that optimizes the SC assignment. Section IV introduces the gradient descent based REINFORCE algorithm. Section V provides system performance evaluation under various scenarios. Finally, the conclusion is given in Section VI.

II. System Model and Problem Formulation

The typical tree-topology architecture of an OFDMA-PON system is presented in Fig. 1, in which three main components are shown, including an optical line terminal (OLT) at the central office (CO), an optical passive splitter-based distribution network (ODN), and many client-end ONUs [33], [34]. The OLT broadcasts the downstream data to each ONU through an ODN. The ODN forwards upstream data from each ONU to the OLT. The ONUs selectively receives downstream frames broadcast by the OLT and transmits them to their client. As specified in IEEE 802.3ca [35], a power-adaptive burst mode receiver equipped at OLT can respond to a large burst input power ratio from the minimum receiver sensitivity (e.g., -28 dBm) to overload (e.g., -6 dBm). Hence, the power difference that is caused by the length difference of distributed fibers to the different ONUs, is assumed to be handled completely by the burst mode receiver, thus it is not considered by this paper.

Generally, the up-/down-stream data traffic are transmitted through an optical wavelength channel, which can be further divided into many OFDM SCs in the frequency domain. In this paper, we focus on the upstream (US) transmission of the OFDMA-PON system, where the total US bandwidth is divided into the many orthogonal SCs, a part of which can be assigned to a different ONU in different TSs. The SCs can be grouped into SC channels, each of which includes one or more SCs. To avoid inter-ONU interference, we assume that a remotely seeded carrier from OLT and an estimation of the differential delay are adopted in each ONU, and thus the carrier frequency offsets (CFO) and frame delay of each US multiple access signal are eliminated significantly [36].

We model the OFDMA-PON system with S SCs and K ONUs, and each SC can only be occupied by one ONU within each TS. The SCs used each time by an ONU must be adjacent. In a TS, each OFDM SC is shown in Fig. 1 as a rectangular block and each SC block stands for a bandwidth of 100 MHz (i.e., $f_{\mathrm{SC}} = 100$ MHz). The finer-granularity SC will improve the spectrum utilization and avoid wasting bandwidth resources. We assume the bandwidth request of each ONU has already included a small portion of guard band (GB) to avoid inter-ONU interference. Note that, since the different modulation formats are allocated to the different SCs according to DSA algorithms for the ONUs, i.e., the SC allocation and bit allocation, the signal transmitting power and traffic average delay of the ONU are different, correspondingly. For example, given an ONU bandwidth request, if the ONU chooses to use a low-order modulation format, the ONU's transmitting power will be lower, but the traffic average delay of the ONU is more likely to be higher because more SC resources would be less available at that time. Note that the modulation format adopted is a key factor that affects system energy consumption and ONU traffic delay performance simultaneously. Therefore, it is important to carefully allocate the TSs, SCs, and the modulation formats for an ONU bandwidth request in OFDMA-PONs.

We assume that the modulation formats of all SCs assigned to one ONU are unique, and hence denote b_k as the number of bits carried by an OFDM symbol for the kth ONU. b_k takes values of $12, \ldots, N$, when the modulation format is from *BPSK* to 2^N-QAM, where N is the maximum number of bits that can be transmitted per OFDM symbol.

As described in [3], since electrical power accounts for a large portion of the ONU's total transmit power, the optical power of the ONU can be ignored. Hence, we design the power consumption model for the ONU as

$$P_k = \frac{E_k}{T_k \cdot a_k^2} \tag{1}$$

where a_k denotes the channel gain of the *kth* ONU, T_k is the duration of the *kth* ONU request in term of the TSs, and E_k represents the required energy consumption to support b_k bits/symbol for the given bit-error-rate (BER) P_e during the T_k. For a multicarrier system transmitting over a linear time-invariant channel with additive white Gaussian noise (AWGN), the required energy consumption E_k can be expressed as [37]

$$E_k = \frac{N_0}{3} \cdot \left[Q^{-1}\left(\frac{P_e}{4}\right)\right]^2 \cdot \left(2^{b_k} - 1\right) \cdot ceil\left(\frac{R_k}{b_k \cdot f_{SC}}\right) \cdot T_k \tag{2}$$

where N_0 is the noise power spectral density; the quality factor is $Q(x) = (1/\sqrt{2\pi}) \int_x^\infty e^{-t^2/2} dt$, P_e is a given BER, R_k is the data rate required by the *kth* ONU. Obviously, given a fixed P_e of 10^{-9}, the value of $(N_0/3) \cdot [Q^{-1}(P_e/4)]^2$is a constant, and equals to 0.4039. Note that Eq. (2) to compute the transmitting power with the basic modulation format (e.g., BPSK) and the corresponding values are also used in [15]. Since the supposition that the signal baud rate is roughly equivalent to the occupied bandwidth by the transmitted signal, the $ceil(R_k/(b_k \cdot f_{SC}))$ is the number of the required SCs by the *kth* ONU. Thus, we simplify Eq. (2) as follows:

$$E_k = 0.4039 \cdot \left(2^{b_k} - 1\right) \cdot ceil\left(\frac{R_k}{b_k \cdot f_{SC}}\right) \cdot T_k \tag{3}$$

Since the number of bits transmitted per OFDM symbol is b_k for the *kth* ONU with the modulation of $2^{b_k}-$QAM and the choice of modulation format would have a major impact on the transmission quality, we need to carefully choose the b_k. Specifically, if the overall signal power remains constant, a larger constellation resulting from a higher b_k value would cause a degraded BER. That is, it requires an increase in signal transmission power to satisfy the BER requirement.

In this paper, our objective is to minimize the average latency and average transmitting power of ONUs by finding an optimal allocation of $b_k (k \in \{1, 2, \ldots, K\})$ on the condition that the traffic demand of each ONU is met.

$$Minimize\left(\alpha \cdot \sum_{k \in K} \frac{c_k - T_k}{T_k} + \beta \cdot \sum_{k \in K} \frac{P_k - P_k^{fix}}{P_k}\right) \tag{4}$$

The α and β are the factors introduced to adjust the weight of the two terms. The first term reflects the normalized total traffic delay, and the second term represents the normalized total ONU transmitting power. c_k is the completion time of the *kth* ONU request. *Pfix k* is the transmitting power of the *kth* ONU with a fixed modulation format capable of satisfying all ONU traffic requirements.

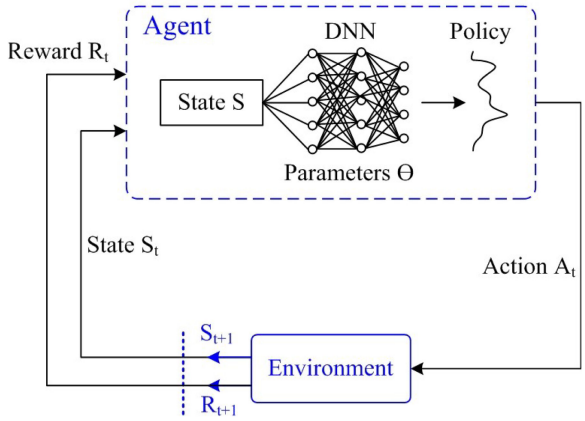

Fig. 2. Demonstrating agent-environment through DNN interaction in RL.

III. Deep Reinforcement Learning Model for Subcarrier Assignment

Fig. 2 shows a typical Markov decision process that learns from agent-environment interactions to achieve certain goals. Learners or decision-makers are called agents. What it interacts with, including everything outside the agent, is called the environment. The agent is constantly interacting with the environment, i.e., the agent chooses an action and then the environment reacts to it and reveals environment changes to the agent. At each time step *t*, the agent observes state S_t and selects an action A_t based on it. As a result of the action, the agent receives an immediate reward R_{t+1} and the state of the environment transfers to S_{t+1}. The goal of the agent is to ultimately achieve higher cumulated rewards or long-term rewards, and often, in order to gain higher cumulated rewards, immediate rewards must be waived.

The proposed DRL-based DSA policy is modeled as a policy network, trained to find the optimal b_k allocation (i.e., the optimal modulation format) that minimizes both the traffic delay and power consumption of ONUs. To establish the DRL-based model for DSA, we define the state, action, and reward of DRL-based policy as follows:

State: We represent the state of the system, which includes the current subcarrier allocation state of the allocated ONU requests (see Fig. 3(a)) and the unallocated ONU requests in request slots (see Fig. 3(b)) and backlog queue (see Fig. 3(c)). Note that Fig. 3(a) is a two-dimensional image. The vertical axis represents the time dimension that begins with the current time step and lasts for *T* steps. The horizontal axis describes the SC resource requirements of the ONU requests. The different colors in the image represent different ONUs. For example, the blue color blocks in Fig. 3(a) indicate that the ONU request which needs two subcarriers and lasts for two time slots is successfully assigned. The Request Slot images in Fig. 3(b) represent the unallocated requests in the case with different options of the modulation formats. For instance, the request in Request Slot 1 has a length of two time slots. Depending on the three different modulation formats, it requires four SCs using BPSK or two SCs using 4-QAM, or one SC using 8-QAM, respectively.

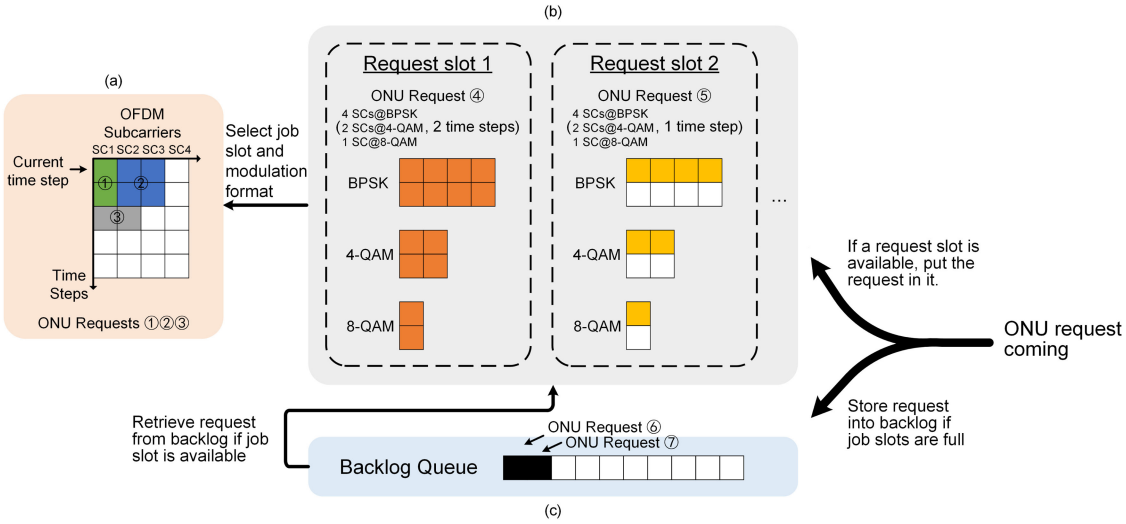

Fig. 3. Example of a state representation with two pending request slots and three modulation formats, (a) Current subcarrier allocation state, (b) Unallocated ONU requests in request slots, and (c) Backlog queue.

Note that, we would prefer to maintain the input of the neural network to be represented in a fixed form images, hence only *M* request slots are set to accommodate the earliest requests to be allocated (e.g., $M = 2$ in Fig. 3(b)) [27]. The information on the remaining ONU requests is stored in a backlog queue as shown in Fig. 3(c). Restricting attention to earlier arriving ONU requests would benefit latency reduction and limit the action space so that the learning process can be more efficient.

Action: For each time step, the agent may want to schedule any subset of the *M* ONU requests. There are *N* alternative modulation formats (i.e., from BPSK to 2^N-QAM), only one of which is selected for scheduling at a time. However, scheduling in this way would require an $(N+1)^M$ action space, which can make the learning process particularly challenging. To address this issue, we divide the actual time step into a few time-frozen steps [27], where the agent can choose only one ONU and designate the modulation format. Once the request is scheduled, it is removed from the corresponding request slot within a time step. In this case, if there is a request queuing in the backlog queue, it will be retrieved and accommodated in the request slot. The agent also needs to determine when to exit the frozen step by adding an exit action into the action space, making the size of the action space being $M \times N+1$. In detail, the action space is given by $\{\emptyset, (1,1), (1,2), \ldots, (1,N), (2,1), \ldots, (M,N)\}$, where $a = (m,n)$ indicates that the modulation format n is selected for the ONU request in request slot m. $a = \emptyset$ means that the agent chooses to quit the frozen step, and no more ONU requests will be scheduled in the current time step. Additionally, at each frozen step, if the remaining resource cannot satisfy the ONU request, the agent is forced to exit the frozen time step as well. Afterward, time will continue rolling (i.e., time step t+1). By this time, the resource pool image, the request slot image, and the backlog queue image will all be updated. During these images are updated, a newly arrived request can be placed in an idle request slot; otherwise, the request is placed in the backlog queue when the request slot is not available.

Reward: We design the reward function to seek the best strategy for our goal, which is to reduce both traffic latency and power consumption by jointly allocating TSs, SCs, and modulation formats. For each time step, we set the reward as

$$R_t = -\alpha \cdot \left(\sum_{j \in J} \frac{1}{T_j} - \sum_{k \in K'} 1 \right) - \beta \cdot \sum_{k \in K'} \frac{P_k - P_k^{fix}}{P_k} \tag{5}$$

where J is the set of ONU requests in the current system, K' is the set of ONU requests scheduled at this time step. T_j is the time length of the *jth* ONU request, i.e., how long the request last. Note that the agent does not receive any intermediate incentive for decisions made in each frozen time step. By interacting with the environment, the agent attempts to select an action to maximize the sum of the discounted rewards it receives in the future. The goal is to maximize the expected cumulative discount rewards: $E[\sum_{t=0}^{\infty} \gamma^t R_t]$, where $\gamma \in (0, 1]$ is the discount rate. By setting $\gamma = 1$, the cumulated reward of the first term $\sum_{j \in J} 1/T_j - \sum_{k \in K'} 1$ coincides with the normalized queuing time of the *jth* ONU request, which consists of the time spent in the backlog queue and the ONU request slot. We use the second term to evaluate the effect of power consumption. α and β are the weights of latency and power consumption. We set the weights $\alpha = \beta = 0.5$, so that the cumulative reward is maximized in order to minimize both delay and transmitting power. At each time step t, the agent observes state S_t and selects an action A_t based on the expected cumulative discount reward. After a time step, the agent receives R_{t+1} and the state of the environment transitions to S_{t+1}.

IV. Gradient Descent Based Reinforce Algorithm

A policy refers to the probability distribution that maps from state space to action space: $\pi(s, a) \rightarrow [0, 1]$. If the agent follows the policy π at time t, then $\pi(a|s)$ is the probability of taking action $A_t = a$ under state $S_t = s$. What RL learns is how to optimize the policy to maximize the expected cumulated reward (also referred to as Return) G. Initially, the policy is stored in a tubular form because there are just a handful of states and

actions. However, with the increasing complexity of problems, the number of state-action pairs is exploding. To handle such an issue, the function approximation approach is introduced. As one of the most popular solutions, DNN is adopted and being used as a policy network. By adding a policy network parameter θ, the policy becomes $\pi(a|s,\theta) = P\{A_t = a|S_t = s, \theta_t = \theta\}$,which represents the probability of taking action *a* under the condition of state being *s* and policy parameter being θ at time *t*. The policy network receives the system state (set of images shown in Fig. 3) as input and the output indicates which ONU request is scheduled and the allocation of b_k.

The policy gradient method, which directly approximates the optimal policy, does not require to form an approximation function. It learns by performing a gradient descent on the policy parameters. The objective is to maximize the expected cumulative reward, and the gradient of the given target is provided by [38]:

$$\begin{aligned}\nabla J(\theta) &= E_\pi\left[\sum\nolimits_a q_\pi(S_t,a)\nabla\pi(a|S_t,\theta)\right]\\ &= E_\pi\left[G_t\frac{\nabla\pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)}\right]\\ &= E_\pi\left[G_t\nabla\ln\pi(A_t|S_t,\theta)\right]\end{aligned} \tag{6}$$

where $q_\pi(s_t,a)$ is the action-value function, defined in Eq. (7). It calculates the expected cumulative reward that selects action *a* under state *s* following policy π.

$$\begin{aligned}q_\pi(s,a) &= E_\pi\left[G_t|S_t = s, A_t = a\right]\\ &= E_\pi\left[\sum_{k=1}^{T-t} R_{t+k}|S_t = s, A_t = a\right]\end{aligned} \tag{7}$$

And G_t is the cumulative reward, the sum of rewards starting from the current time step *t* to the terminal step *T*. It is defined by:

$$G_t = \sum_{k=1}^{T-t} R_{t+k} \tag{8}$$

The key idea of the policy gradient method is to estimate the gradient by observing the execution trajectory: (S_0, A_0, R_1, S_1, A_1, R_2, …, $S_{T\text{-}1}$, $A_{T\text{-}1}$, R_T) following the policy π_θ, then update the policy network parameters θ by gradient descent as follows:

$$\theta_{t+1} = \theta_t + \alpha G_t\nabla\ln\pi(A_t|S_t,\theta) \tag{9}$$

where α is the step size. This algorithm is also known as the REINFORCE algorithm [38].

The increment is proportional to $\nabla\ln\pi(A_t|S_t,\theta)$ and G_t, where $\nabla\ln\pi(A_t|S_t,\theta)$ is the direction that increases the probability of picking action A_t and the return G_t indicates how far should the θ update is going in this direction, i.e., if G_t is positive and large, then the corresponding action A_t is preferred and vice versa. This allows the policy network parameters to move toward the direction that favors the action that yields the highest return.

We train the policy network in an episodic manner. In each epoch, a fixed number of ONU requests arrive and are scheduled according to the policy. Each epoch ends with all ONU requests being scheduled. Table I shows the pseudo-code for the training algorithm.

TABLE I: Pseudocode for Training Algorithm.

Input: Differentiable parameterization policy $\pi(A_t|S_t,\theta)$
Output: Optimal resource allocation policy
1: Initialization of policy parameters: $\theta \leftarrow 0$;
2: **For** each ONU request set:
3: **For** each episode *i = 1, …, L*:
4: **For** each action in episode *t = 0, 1, …, T-1*:
5: generate $S_0^i, A_0^i, R_1^i, \ldots, S_{T-1}^i, A_{T-1}^i, R_T^i$, follow policy $\pi(\cdot|\cdot,\theta)$
6: $G^i \leftarrow \sum_{k=t+1}^{T} R_k^i$
7: $b^i \leftarrow \frac{1}{N}\sum_{i=1}^{N} G^i$
8: $\theta \leftarrow \theta + \alpha(G^i - b^i)\nabla\ln\pi(A_t|S_t,\theta)$
9: **End**
10: **End**
11: **End**

In order to train a generalized policy, we use multiple sets of ONU requests as the training set, which we refer to as the ONU request set. In each training iteration, we repeatedly run *L* rounds of all *C* ONU request sets to explore the probability space of possible actions using the current policy and use the resulting data to improve the scheduling policy. Specifically, a total of *C·L* trajectories are recorded, including the state, action, and reward for each step and these data are used to calculate the discounted cumulative reward G_t for each time step *t* of each episode. We then train the neural network using a modified version of the previously described REINFORCE algorithm.

Equation (9) is used in the original REINFORCE algorithm to estimate the policy gradient. The downside of this equation is that the gradient estimate has a high variance, which can be reduced by subtracting a baseline b_t. The baseline b_t is the average of return G_t (see line 7 in Table I). The θ update equation is then rewritten as follows:

$$\theta_{t+1} = \theta_t + \alpha\left(G_t - b_t\right)\nabla\ln\pi\left(A_t|S_t,\theta\right) \tag{10}$$

V. Performance Evaluation

A. Simulation Setup

In our simulations, the ONU requests arrive according to the Bernoulli process. The request arrival rate λ (i.e., a new ONU request arrives with a certain probability at each time step) ranges from 0 to 1 with a step size of 0.1. There are 128 SCs for the upstream transmission and the total bandwidth is 12.8 GHz. To reduce computational complexity, only 32 SC channels are considered, and each channel includes 4 SCs. All ONUs support upstream SC channels with the same transmission characteristics. The duration of ONU requests is set as follows: 80% of the ONU requests have a time length chosen uniformly between 1t and 3t; the remainder is chosen uniformly from 10t to 15t.

The SC modulation format can be chosen from BPSK, 4-QAM, 8-QAM, and 16-QAM. Our proposed DRL-based scheme can flexibly select the most suitable modulation format according to the different bandwidth demands of the ONUs, which is compared against four benchmarks using a fixed modulation format. In the benchmarks, the lowest-order modulation format would be adopted for all ONU requests, even though the ONUs all have very different bandwidth requirements. It is because that using a fixed lowest-order modulation format is an effective way to reduce the required power consumption to send ONUs' signals.

We assume that the bandwidth requirement of each ONU is uniformly distributed in the range of $[\mathrm{D}_{\min}, \mathrm{D}_{\max}]$, and the lower bound $\mathrm{D}_{\min}$ is set to 3.2 Gb/s.

With the different values of the upper bound $\mathrm{D}_{\max}$, the following three cases are considered:

- *Case 1:* when the $\mathrm{D}_{\max} \in (12.825.6]$ Gb/s, the fixed modulation format is 4-QAM for other benchmarks. In this case, without loss of generality, we set $\mathrm{D}_{\max}$ to a median value as (12.8+25.6)/2 = 19.2 Gb/s. Thus, the range of values for bandwidth demand of each ONU is [3.2, 19.2] Gb/s, which corresponds to [16, 96] SCs modulated with 4-QAM.
- *Case 2:* when the $\mathrm{D}_{\max} \in (25.638.4]$ Gb/s, the fixed modulation format is 8-QAM for other benchmarks. Likewise, we set $\mathrm{D}_{\max}$ to a median value as 32 Gb/s. Thus, the range of values is [3.2, 32] Gb/s, corresponding to [8, 96] SCs modulated with 8-QAM.
- *Case 3:* when $\mathrm{D}_{\max} \in (38.451.2]$ Gb/s, other benchmarks would adopt the 16-QAM as the fixed modulation format. So we can set $\mathrm{D}_{\max}$ to a median value as 44.8 Gb/s. Thus, the range of values is [3.2, 44.8] Gb/s, corresponding to [4, 96] SCs modulated with 16-QAM.

Hence, those ONU whose bandwidth demand larger than 51.2 Gb/s cannot be satisfied, even with the highest-order modulation format available in the system. Note that in each case, the ONU bandwidth request has already included a two-SC GB between every two SC groups coming from different ONUs, in order to avoid the inter-ONU interference. For instance, if the ONU bandwidth requirement is 3.2Gb/s, 4 SC channels (i.e., 16 SCs in total) are assigned to the ONU with 4-QAM modulation, where 14 SCs are used for US data transmission, and the other 2 SCs are reserved as guard band.

The four benchmarks are: 1) Random, which selects ONU requests randomly, 2) the shortest request first algorithm (SRF), which serves ONU requests in the ascending order of their duration [30], 3) the resource wrapper Packer algorithm [31], which assigns resource according to the order of alignment between resource requirements and resource availability, 4) the synthesis Tetris algorithm, which balances the advantages of taking short-term request and resource packaging [32].

In this simulation, the time window of the State space observed by the DRL agent is 20t long and each scheduling episode for each ONU request set lasts 50t. The agent picks a serving ONU request from $M = 8$ ONU request slots and chooses a suitable modulation format from $N = 4$ candidate modulation format, while observing other ONU requests in the backlog queue. The length of the backlog is set to 64. As mentioned in the Section III, part "Action", the size of action space is $M \times N + 1 = 33$, and hence the output layer of DNN has 33 neurons. Because the hidden layer is fully connected to the output layer, the number of neurons on the hidden layer should be the multiples of 33 neurons. After multiple testing, we found that a hidden layer of 33 neurons achieves the best performance. Therefore, the policy network is realized by a DNN with a fully connected hidden layer of 33 neurons and a total of 532323 parameters. The activation function used for the DNN is Rectified Linear Unit (ReLU) [38]. For each training iteration, we use 50 ONU request sets and run 10 Monte Carlo simulations in parallel for each request set. We update the policy network parameters with a learning rate of 0.001.

B. Simulation Results and Discussions

The simulation results drawn below has been processed. In our optimization objective as shown in Eq. (4), the two targets under investigation (i.e., average traffic latency and average transit power) vary in the different range of values, which might even be not the same order of magnitude (refer to Fig. 10). To obtain the same level of optimization for the two targets in the DRL-based DSA policy, we apply a balance coefficient ρ to all data of the average traffic latency. The ρ value can be determined when both of the two targets achieve optimal states in the case of $\alpha = \beta = 0.5$. Having run numerous tests, we set ρ value to be 0.087, 0.124 and 1, respectively, in the Case 1, Case 2, and Case 3. It is due to the fact that the optimization scopes of the two targets are also not identical in the different cases of the ONU bandwidth demands. In the following simulations, the two weights are set the same (i. e., $\alpha = \beta = 0.5$), to equally treat the two above target metrics.

In Fig. 4, test results of total reward, average traffic latency, and average transmit power are given out in Case 1 with $R_{\mathrm{k}} \in$ [0.32,1.92] Gb/s, where the fixed modulation format of the four benchmarks is set to be 4-QAM by the above rules. Fig. 4(a) and (b) compare the total reward and the traffic delay performance under the variance of the ONU request arrival rate. Compared to the benchmarks Packer and Random, SRF has better performance in terms of both reward and traffic delay. In light-load conditions, SRF performs similar to Packer. As load increases, the performance gap between SRF and Packer continues to increase and SRF is approaching Tetris. While Packer reserves more resources for large requests than SRF, more requests are using reserved resources under heavy load, which directly brings Packer to the worst delay performance. The Tetris outperforms the SRF and Packer by combing their advantages. As shown, the proposed DRL performs better than heuristics for the three metrics under high load conditions due to the ability of the DRL to learn to assign modulation formats flexibly to ONU requests with different bandwidth requirements to save power (Fig. 4(c)) and to reserve resources for potential low demand requests to reduce latency (Fig. 4(b)), while the total reward is maximized.

Figs. 5 and 6 demonstrate test results in Case 2 with $R_{\mathrm{k}} \in$ [0.32, 3.2] Gb/s, and in Case 3 with $R_{\mathrm{k}} \in$ [0.32, 4.48] Gb/s, respectively. In the Case 2 and Case 3 of the ONU bandwidth requirements, the fixed modulation format of the benchmarks is set to be 8-QAM and 16-QAM, respectively. From these figures,

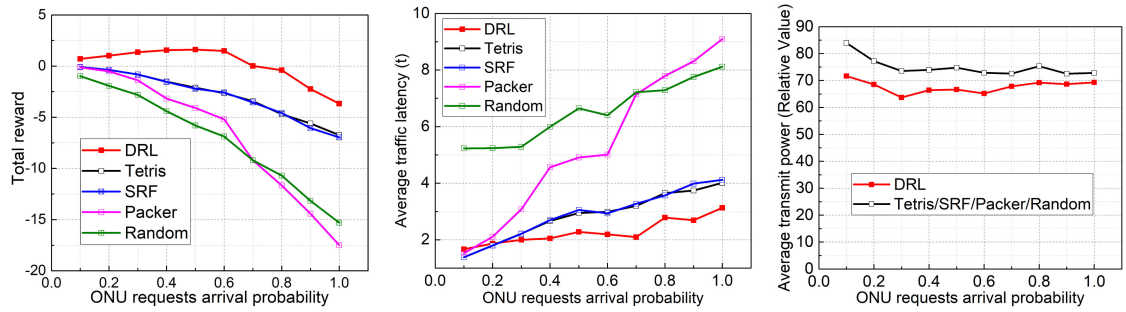

Fig. 4. Test results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 1 with $R_k \in [3.219.2]$ Gb/s, when α=β=0.5.

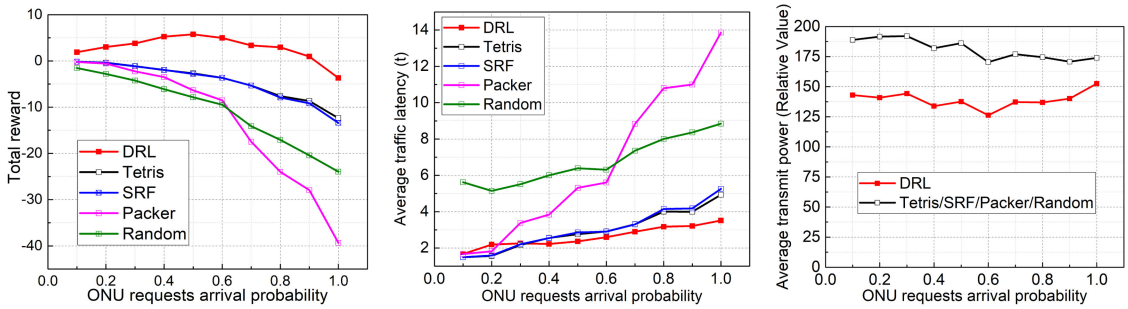

Fig. 5. Test results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 2 with $R_k \in [3.232]$ Gb/s, when α=β=0.5.

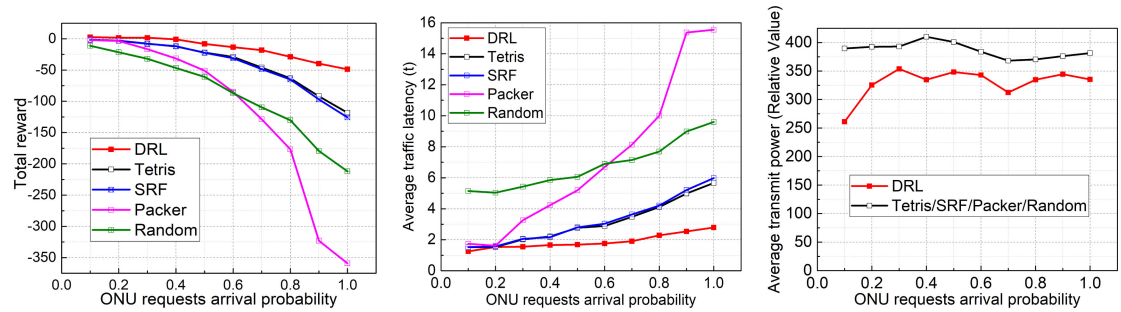

Fig. 6. Test results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 3 with $R_k \in [3.244.8]$ Gb/s, when α=β=0.5.

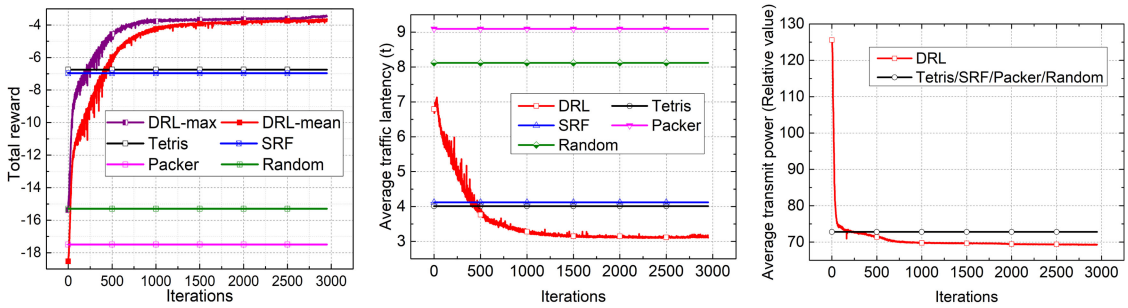

Fig. 7. Training results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 1 with $R_k \in [3.219.2]$ Gb/s, when α=β=0.5 and λ=1.

we can find that the three performance metrics of the proposed DRL is better than that of all the benchmark heuristics. In addition, from Case 1 to Case 3, as the ONU bandwidth requirements increase, the superiority of the proposed DRL approach become clearer, comparing with other benchmark heuristics. This is because the increase in the ONU bandwidth demands makes the DRL allocate modulation format more flexibly, which does not only save the power, but also reduce the traffic delay. Moreover, the reduction in power consumption contributes more to the total reward.

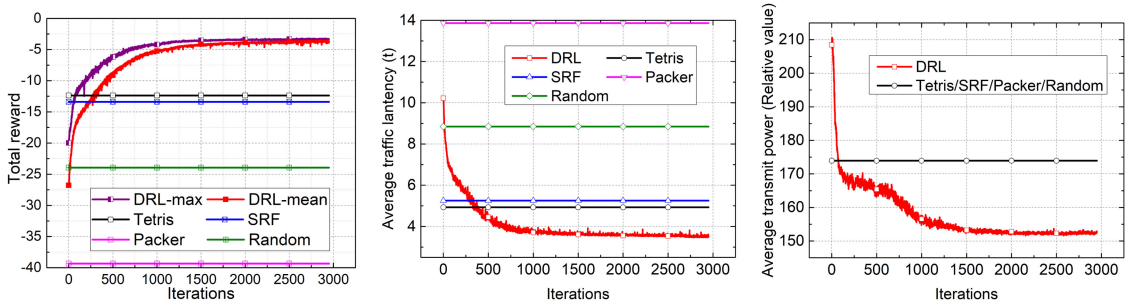

Fig. 8. Training results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 2 with $R_k \in [3.2, 32]$ Gb/s, when $\alpha=\beta=0.5$ and $\lambda=1$.

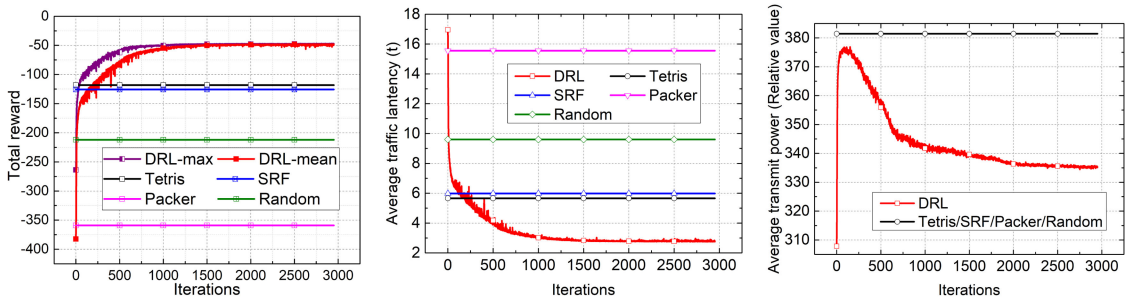

Fig. 9. Training results of (a) Total reward, (b) Average traffic latency, (c) Average transmit power in Case 3 with $R_k \in [3.244.8]$ Gb/s, when $\alpha=\beta=0.5$ and $\lambda=1$.

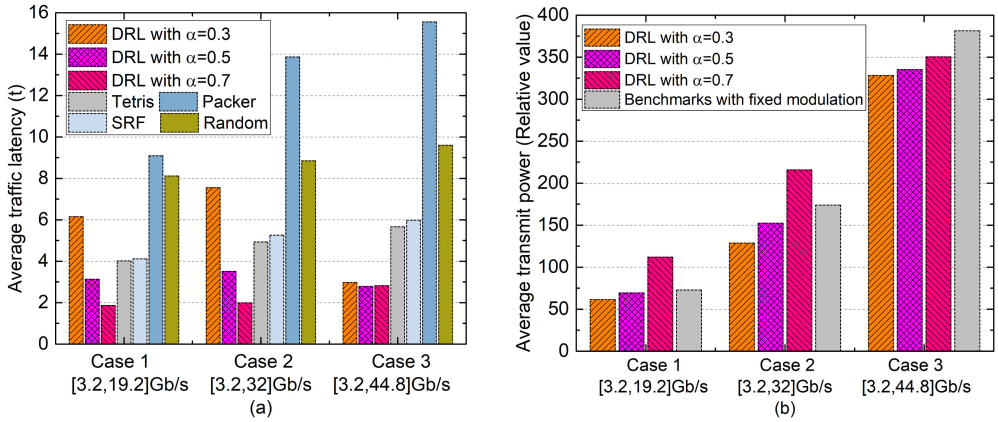

Fig. 10. The impact of weight coefficient α on the test results of (a) Average traffic latency, (b) Average transmit power, in the different cases of R_k.

Fig. 7 demonstrates how the DRL agent learns over the training iterations in Case 1 with $R_k \in [0.32,1.92]$ Gb/s, when the request arrival rate λ is 1. As stated in Section IV, the cumulative reward (i.e., total reward) of each trajectory is G_0, which can be calculated by Eq. (8). Hence, for all $C{\cdot}L$ training trajectories, there are $C{\cdot}L$ G_0 (i.e., G_1^0, G_2^0, …, $G_{C\cdot L}^0$). The max(G_1^0, G_2^0, …, $G_{C\cdot L}^0$) of all $C{\cdot}L$ trajectories recorded in a training iteration is $\mathrm{DRL}_{\max}$, while the mean value G_0 of all $C{\cdot}L$ trajectories in a training iteration is defined as $\mathrm{DRL}_{\mathrm{mean}}$. At the beginning of the iteration, the DRL does not have any prior knowledge about the dynamics of the system. Its behavior is therefore similar to a random strategy and behaves worse than all the benchmarks. As the iteration continues, both the $\mathrm{DRL}_{\max}$ and $\mathrm{DRL}_{\mathrm{mean}}$ are steadily increased with continuous training. After about 100 training iterations, the DRL agent learns that it can improve the total reward by reserving some resources for small requests and using the lower modulation formats more frequently. Afterward, the DRL proceeds to try to increase the total reward until the difference between $\mathrm{DRL}_{\max}$ and $\mathrm{DRL}_{\mathrm{mean}}$ gradually converges to a stable value after 1500 iterations, implying that the system has reached an optimum state. The simulation results in Fig. 7(b) and (c) show that the DRL-based scheme achieves the optimal traffic delay while at the same time reducing the transmission power as much as possible.

Figs. 8 and 9 show the training results in Case 2 with $R_k \in [0.32, 3.2]$ Gb/s and Case 3 with $R_k \in [0.32, 4.48]$ Gb/s, respectively, when the λ is 1. Compared with the training results of Case 1 in Fig. 7, the DRL still learns to reserve some resources

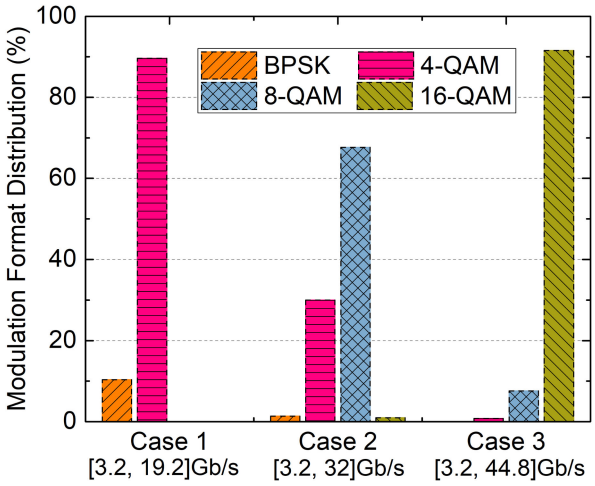

Fig. 11. The distribution of the allocated modulation formats in the different cases of R_{k}, when using the proposed DRL method and $\lambda = 1$.

for small requests at first and tends to use lower modulation formats to obtain a higher reward afterward. After about 1500 iterations, the system reaches its optimum state, where all three indicators are optimal. Note that in Fig. 9(c), although DRL continues to attempt to further increase the value of the total reward by slightly adjusting the modulation format to reduce the transmission power, there is no significant improvement which indicates that the system has reached its optimum state at this point.

In Fig. 10, we investigate the impact of the different α and β weight values on the following performance metrics: (a) service delay and (b) transmit power, when λ is set to 1. In Fig. 10(a), when $\alpha = 0.7$, the DRL achieves the lowest service latency, since the DRL attaches more attention to the traffic latency rather than the transmit power. In the case, the DRL would tend to select a higher modulation format to reduce service latency. In Fig. 10(b), the DRL with $\alpha = 0.3$ (i.e., $\beta = 0.7$) has the lowest transmission power, because the DRL prefers to optimize the transmit power. These above results indicate that the importance of the two metrics can be adjusted by changing the values of α and β, and that the DRL can be specially trained to optimize for customized objectives. In addition, we notice that compared with the DRL with the $\alpha = \beta = 0.5$, which represents a baseline scenario, the benchmark algorithms face higher traffic delay and transmission power as the traffic load increases. However, the DRL can adapt to all kinds of load conditions and achieve optimal results. This is because, as load increases, the flexible and adaptive assignment of the modulation formats for ONU bandwidth demands would facilitate to use more efficiently these network resources.

Fig. 11 shows the distribution of the allocated modulation formats in different cases of R_k, when using the proposed DRL method and the λ is 1. It is observed that after iterative training of the DRL, coincidentally the most frequently used modulation format is just what the benchmarks adopts, i.e., 4-QAM for Case 1, 8-QAM for Case 2 and 16-QAM for Case 3. As mentioned above, the fixed modulation format adopted by the benchmarks depends on the upper bound $\mathrm{D}_{\max}$ of the ONU bandwidth demand, since the fixed modulation format is required to be applied for all the ONUs. Hence, the available lowest-order modulation format is chosen, and it can effectively reduce the required power consumption to send ONUs' signals. In addition, to further reduce the power consumption, a much smaller set of the ONU requests with the smaller bandwidth demand can be assigned with the lower modulation format than the most frequently used one.

VI. Conclusion

We have proposed a novel three-dimensional DRL-based DSA algorithm in OFDMA-PON, which jointly allocates the TSs, SCs, and modulation formats to optimize the average delay and average power consumption of the ONU requests simultaneously. Simulation results show that the proposed DRL-based DSA scheme can significantly reduce average delay and average power consumption, compared to benchmark scheduling schemes such as the SRF, Packer, Tetris, Random strategies. From the analysis of the simulation results, we found that the reason for DRL to be able to achieve better performance is that a more flexible and adaptive modulation scheme is adopted. Since the DRL agent can improve itself by directly learning from experience, it is a powerful and versatile tool for many optimization issues in future optical networks.

References

[1] K. Kanonakis, E. Giacoumidis, and I. Tomkos, "Physical-Layer-Aware MAC schemes for dynamic subcarrier assignment in OFDMA-PON networks," *J. Lightw. Technol.*, vol. 30, no. 12, pp. 1915–1923, Jun. 2012.
[2] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT a dynamic protocol for an ethernet PON (EPON)," *IEEE Commun. Mag.*, vol. 40, no. 2, pp. 74–80, Feb. 2002.
[3] W. You, L. Yi, S. Huang, J. Chen, and W. Hu, "Power efficient dynamic bandwidth allocation algorithm in OFDMA-PONs," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 5, no. 12, pp. 1353–1360, Dec. 2013.
[4] J. Zhang *et al.*, "A clock-gating-based energy-efficient scheme for ONUs in real-time IMDD OFDM-PONs," *J. Lightw. Technol.*, vol. 38, no. 14, pp. 3573–3583, Jul. 2020.
[5] W. Lim *et al.*, "Dynamic bandwidth allocation for OFDMA-PONs using hidden Markov model," *IEEE Access*, vol. 5, pp. 21016–21019, 2017.
[6] K. Fouli and M. Maier, "OCDMA and optical coding: Principles, applications, and challenges [Topics in optical communications]," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 27–34, Aug. 2007.
[7] E. Wong, "Next-Generation broadband access networks and technologies," *J. Lightw. Technol.*, vol. 30, no. 4, pp. 597–608, Feb. 2012.
[8] W. Jin *et al.*, "Hybrid SSB OFDM-Digital filter multiple access PONs," *J. Lightw. Technol.*, vol. 38, no. 8, pp. 2095–2105, Apr. 2020.
[9] J. Ma *et al.*, "Cost-effective SFO compensation scheme based on TSs for OFDM-PON," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 11, no. 6, pp. 299–306, Jun. 2019.
[10] M. Bi, S. Xiao, and W. Li, "Joint subcarrier channel and time slots allocation algorithm in OFDMA passive optical networks," *Opt. Commun.*, vol. 287, pp. 90–95, 2013.
[11] W. Lim, K. Kanonakis, P. Kourtessis, M. Milosavljevic, I. Tomkos, and J. M. Senior, "Flexible QoS differentiation in converged OFDMA-PON and LTE networks," in *Proc. Opt. Fiber Commun. Conf. Expo. Nat. Fiber Opt. Engineers Conf.*, 2013, pp. 1–3.
[12] J. Zhang *et al.*, "Decision-Feedback frequency-domain volterra nonlinear equalizer for IM/DD OFDM long-reach PON," *J. Lightw. Technol.*, vol. 37, no. 13, pp. 3333–3342, Jul. 2019.
[13] Y. Li, J. Han, and X. Zhao, "Performance investigation of a Cost- and Power-Effective high nonlinearity tolerance OFDMA-PON scheme based on sub-nyquist sampling rate and DFT-Spread," *IEEE Access*, vol. 7, pp. 43137–43142, 2019.

[14] Y. Li, J. Han, L. Han, and C. Ju, "Investigation of a flexible downstream scheme for sampling rate and bandwidth reduction in short reach communication systems," *IEEE Access*, vol. 8, pp. 59083–59090, 2020.
[15] X. Gong, L. Guo, and Q. Zhang, "Joint resource allocation and software-based reconfiguration for energy-efficient OFDMA-PONs," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no. 8, pp. C75–C85, 2018.
[16] M. Kalil, A. Moubayed, A. Shami, and A. Al-Dweik, "Efficient low-complexity scheduler for wireless resource virtualization," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 56–59, Feb. 2016.
[17] X. Hu, L. Zhang, P. Cao, K. Wang, and Y. Su, "Energy-efficient WDM-OFDM-PON employing shared OFDM modulation modules in optical line terminal," *Opt. Exp.*, vol. 20, no. 7, pp. 8071–8077, 2012.
[18] W. You, "Randomized dynamic bandwidth allocation algorithm for upstream access in OFDMA-PON," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 7, no. 6, pp. 597–601, 2015.
[19] C. Kim, S. Jung, and S. Han, "Microwave photonic filter based all-optical virtual private network supporting dynamic bandwidth allocation in OFDMA-PON system," *IEEE Photon. J.*, vol. 10, no. 1, Feb. 2018, Art. no. 7900208.
[20] H. Bang, K. Doo, S. Myong, G. Stea, and C. Park, "Design and analysis of IPACT-based bandwidth allocation for delay guarantee in OFDMA-PON," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 5, no. 11, pp. 1236–1249, Nov. 2013.
[21] Y. Senoo *et al.*, "Fairness-aware dynamic sub-carrier allocation in distance-adaptive modulation OFDMA-PON for elastic lambda aggregation networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 7, pp. 616–624, Jul. 2017.
[22] W. Lim, P. Kourtessis, K. Kanonakis, M. Milosavljevic, I. Tomkos, and J. M. Senior, "Dynamic bandwidth allocation in heterogeneous OFDMA-PONs featuring intelligent LTE-A traffic queuing," *J. Lightw. Technol.*, vol. 32, no. 10, pp. 1877–1885, May 2014.
[23] X. Hu *et al.*, "High-capacity and low-cost long-reach OFDMA PON based on distance-adaptive bandwidth allocation," *Opt. Exp.*, vol. 23, no. 2, pp. 1249–1255, 2015.
[24] Y. Lv, N. Jiang, D. Liu, C. Xue, and K. Qiu, "Energy-Efficient scheme based on sub-band grouping and allocating for digital filter multiple access adopted PON," *IEEE Photon. J.*, vol. 9, no. 3, Jun. 2017, Art. no. 7904009.
[25] Z. Zhang, Y. Zhu, W. Zhu, H. Chen, X. Hong, and J. Chen, "Iterative point-wise reinforcement learning for highly accurate indoor visible light positioning," *Opt. Exp.*, vol. 27, no. 16, pp. 22161–22172, 2019.
[26] M. Raza, C. Natalino, P. Öhlen, L. Wosinska, and P. Monti, "Reinforcement learning for slicing in a 5G flexible RAN," *J. Lightw. Technol.*, vol. 37, no. 20, pp. 5161–5169, Oct. 2019.
[27] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. 15th ACM Workshop Hot Top. Netw.*, 2016, pp. 50–56.
[28] Z. Gao, J. Zhang, S. Yan, Y. Xiao, D. Simeonidou, and Y. Ji, "Deep reinforcement learning for BBU placement and routing in C-RAN," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2019, pp. 1–3.
[29] X. Luo, C. Shi, L. Wang, X. Chen, Y. Li, and T. Yang, "Leveraging double-agent-based deep reinforcement learning to global optimization of elastic optical networks with enhanced survivability," *Opt. Exp.*, vol. 27, no. 6, pp. 7896–7911, 2019.
[30] H. Mao, M. Schwarzkopf, S. B. Venkatakrishnan, Z. Meng, and M. Alizadeh, "Learning scheduling algorithms for data processing clusters," in *Proc. ACM Special Int. Group Data Commun.*, Beijing, China, 2019, pp. 270–288.
[31] H. Meng, D. Chao, R. Huo, Q. Guo, X. Li, and T. Huang, "Deep reinforcement learning based delay-sensitive task scheduling and resource management algorithm for multi-user mobile-edge computing systems," in *Proc. 4th Int. Conf. Math. Artif. Intell.*, 2019, pp. 66–70.
[32] Y. Bao, Y. Peng, and C. Wu, "Deep Learning-based job placement in distributed machine learning clusters," in *Proc. IEEE Conf. Comput. Commun.*, Paris, France, 2019, pp. 505–513.
[33] B. Chen, M. Zhu, J. Gu, T. Shen, X. Ren, and C. Shi, "Deep reinforcement learning based policy for power efficient dynamic subcarrier assignment in OFDMA-PONs," in *Proc. Asia Commun. Photon. Conf.*, 2019, pp. 1–3.
[34] X. Xue, W. Ji, K. Huang, X. Li, and S. Zhang, "Tunable multiwavelength optical comb enabled WDM-OFDM-PON with source-free ONUs," *IEEE Photon. J.*, vol. 10, no. 3, Jun. 2018, Art. no. 7202008.
[35] *IEEE Standard for Ethernet Amendment 9: Physical Layer Specifications and Management Parameters for 25 Gb/s and 50 Gb/s Passive Optical Networks*, IEEE Standard 802.3ca-2020, pp. 1–267, Jul. 2020.
[36] C. Ruprecht *et al.*, "37.5 km urban field trial of OFDMA-PON using colorless ONUs with dynamic bandwidth allocation and TCM [invited]," *J. Opt. Commun. Netw.*, vol. 7, no. 1, pp. A153–A161, Jan. 2015.
[37] J. G. Proakis, *Digital Communications*, 5th ed. New York, NY, USA: McGraw-Hill, 2008.
[38] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.