# Deep Reinforcement Learning Based Policy for Power Efficient Dynamic Subcarrier Assignment in OFDMA-PONs

**Bin Chen [2], Min Zhu [1], Jiahua Gu [1], Tianyu Shen [2], Xueqi Ren [2], Chenglin Shi [2]**
*1 National Mobile Communications Research Laboratory, Southeast University, Nanjing, 210096, China*
*2 School of Electronic Science and Engineering, Southeast University, Nanjing, 210096, China*
*E-mail address: minzhu@seu.edu.cn*

**Abstract:** The paper proposes a deep reinforcement learning (DRL) based policy for power efficient dynamic subcarrier assignment in OFDMA-PONs. The simulation results show it can reaches the near-optimal traffic delay with a significant power saving.
**OCIS codes:** (060.4510) Optical communications, (060.4256) Networks, network optimization.

## 1. Introduction

As a cost-effective optical access, Orthogonal Frequency Multiplexing Access Passive Optical Networks (OFDMA-PONs) can ensure high capacity and flexible granularity of bandwidth allocation, thus provides an opportunity for optical convergence of 5G ubiquitous wireless access connectivity [1]. Since OFDMA-PONs allow different optical network units (ONUs) to occupy multiple subcarriers (SCs) at the same time on-demand, an appropriate dynamic SC assignment (DSA) strategy for OFDMA-PONs is required to reduce required transmission power and strict traffic latency [2]. The strategy needs to handle the following challenges: 1) seek for the appropriate schedule order for multiple ONU requests to reduce the traffic latency; 2) find the optimal allocation of modulation format in SCs to reduce the transmission power. Importantly, different modulation scheme allocation among SCs will influence both the transmitting power and traffic delay of ONUs. When the BER remains constant, a higher modulation format requires an increase in signal transmitting power. Rather, if the ONU selects SC with low modulation scheme, the overall traffic delay would be higher because the ONU needs more SCs, although its transmitting power will be lower. Therefore, to realize power efficient DSA in OFDMA-PONs, it is essential to allocate time slot, SC, and modulation format jointly, since both the transmitting power and traffic delay of ONUs are closely correlated and we need to balance the utilizations of time slot, SC, and modulation format carefully.

Recently, deep reinforcement learning (DRL) has been applied successfully to some complicated decision problems of resource management. In particular, it can model complex systems and decision strategies as deep neural networks, achieving optimal mapping from state space into action space. In [3], packaging tasks with multiple resource requirements is translated into learning problems to minimize work delays. In [4], a slice admission strategy based on deep reinforcement learning is studied to maximize the profit of infrastructure providers. In [5], a joint BBU placement and routing strategy in C-RAN based on DRL is proposed to achieve better performance.

Inspired by this, we propose a power efficient DSA policy based on deep reinforcement learning in OFDMA-PONs to reduce the traffic latency and transmitting power. The simulation results show that the performance of the proposed DRL-based scheme far exceeds benchmark algorithms, in terms of the latency and the transmission power.

## 2. Problem formulation

Figure1 presents the typical tree-topology architecture of an OFDMA-PON system. The optical line terminal (OLT) broadcasts the downstream service data stream to each ONU through an optical splitter-based distribution network (ODN). The ODN collects data from each upstream ONU and transmits it to the OLT. The up-/down-stream channel can be divided into OFDM SCs, each of which can be allocated to different ONUs in different time slots.

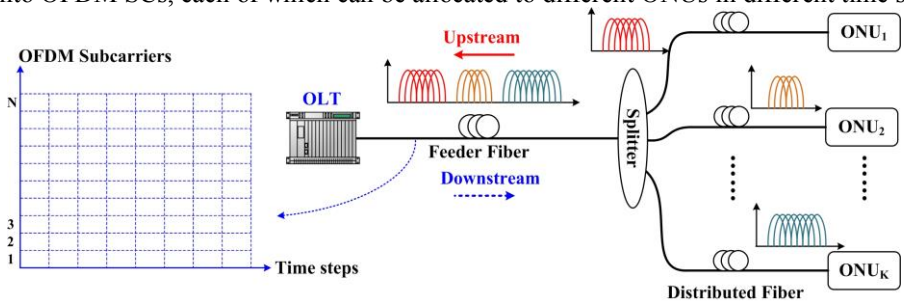


Fig.1. OFDMA-PON architecture

We model the OFDMA-PON system has *N* SCs and *K* ONUs, and each SC can only be occupied by one ONU within each time slot. The SCs used by an ONU are adjacent, and all their start/end times are equal. We assume the modulation format of all SCs occupied by an ONU is the unique. We define $b_k$ as the bits allocated to SCs of the *kth* ONU (i.e., the corresponding modulation format is $2^{b_k}$ QAM).

We design the *kth* ONU power consumption model as $P_k = (N_0/3)\times[Q^{-1}(P_e/4)]^2 \times(2^{(b_k)}-1)\times T_k \times ceil(R_k/b_k)$, where $P_k$ denotes the required ONU transmitting power for supporting $b_k$ bits/symbol for a given bit-error-rate (BER) $P_e$. $N_0$ is the noise power spectral density, the quality factor $Q(x) = \left(1/\sqrt{2\pi}\right)\int_x^\infty e^{\left(-t^2/2\right)}dt$, $T_k$ is the duration of the *kth* ONU request, $R_k$ is the data rate request of the *kth* ONU, the value of $(N_0/3)\times[Q^{-1}(P_e/4)]^2$ is equal to 0.4039 [2]. Thus, we simplify $P_k$ as $P_k = 0.4039\times(2^{(b_k)}-1)\times T_k \times ceil(R_k/b_k)$.

The objective is $Minimize\left(\alpha\times\sum_{k\in K}\left(C_k/T_k\right)+\beta\times\sum_{k\in K}\left(P_k/P_k^{fix}\right)\right)$, to minimize the latency and transmission power jointly. The α and β are the factors introduced to adjust the importance of the two terms, which should be set based on the actual situation. The first term reflects the normalized total traffic delay. The second term represents the normalized total ONU transmitting power, where $C_k$ is the completion time of the *kth* ONU request, $P_k^{fix}$ is the transmitting power of the *kth* ONU with a fixed modulation format, which can cater to all ONU traffic requirements.

## 3. DRL-based Model for DSA in OFDMA-PON

The proposed DRL strategy is modeled as a policy network, trained to find the optimal $b_k$ allocation (i.e., optimal modulation format) to minimize 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 system state as different images. The leftmost image in Fig.2 indicates the allocated ONU requests and start from the current time step and lasting for *T* steps until all ONU requests end. The different colors in these images represent different ONUs. The Request Slot images represent the resource requirements of the ONU requests to be allocated with different modulations. We would like to maintain images for only *M* requests that arrive first so that the input of the neural network can be represented as a fixed state representation (*M*=2 in Fig.2). The other ONU requests to be allocated are stored in the backlog queue.

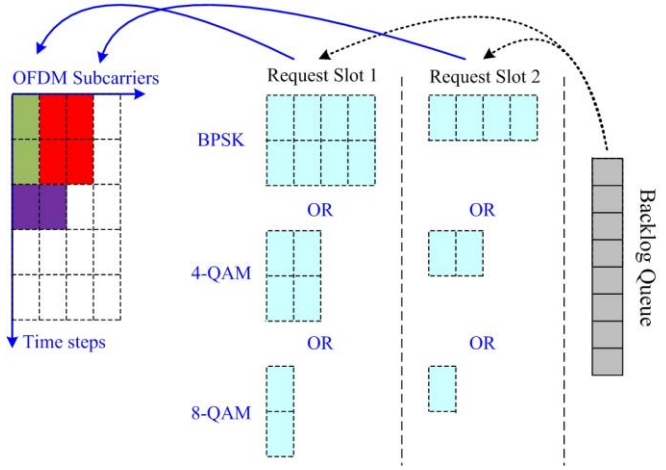


Fig.2. An example of a state representation with two pending request slots and 3 modulation formats

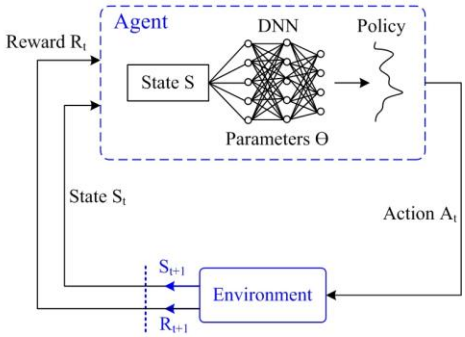


Fig.3. Demonstrating agent-environment interaction in RL through deep neural networks (DNN)

**Action**: In Fig. 3, we allow the agent to schedule multiple ONU requests at the same time. The agent select only one modulation format for the ONU request in *Request Slot-i* to move to an appropriate position in the SC resource image; when the agent selects an invalid action or the current resource cannot satisfy the ONU request, the time step moves forward one step and the SC resource image moves up one step. The newly arrived ONU request will inform the agent and update it at the same time.

**Reward:** We develop rewards to seek for best strategies for our goals. Our goal is to reduce both traffic latency and power consumption by jointly allocating time slots, SCs, and modulation format. For each time step, we set the reward $-\alpha\times\sum_{j\in J}\left(1/T_j\right)-\beta\times\sum_{k\in K'}\left(P_k/P_k^{fix}\right)$, where *J* is the set of ONU requests in the current system, $K'$ is the set of the ONU requests to be scheduling at that time step. At each time step *t*, the agent observes some states $S_t$ and

selects an action $A_t$ based on this reward $R_t$. After a time step, the agent receives a new $R_{t+1}$ and the state of the environment transitions to $S_{t+1}$.

By interacting with the environment, the agent attempts to select an action to maximize the sum of the discount rewards it receives in the future. The expected cumulative discount rewards: $E\left[\sum_{t=0}^{\infty}\gamma^t R_t\right], \gamma \in (0,1]$ is the discount rate. Deep neural networks (DNNs) have been used as function approximations to solve tasks successfully. The DNN receives the system state as input. The output of the DNN indicates which ONU requests schedule first and the optimal allocation of $b_k$. We set the discount rate $\gamma = 1$, so we maximize the cumulative reward to minimize both the delay and transmitting power. At each training iteration, the DNN policy network is optimized using the gradient descent method [3].

## 4. Simulation setup and discussion

The ONU requests arrive according to the Bernoulli process. We change the request arrival rate λ from 0 to 1 with a step size of 0.1. To reduce computational complexity, we only consider 32 SC channels, and each channel includes 4 SCs. Thus, there are 128 SCs in the system and the total bandwidth is 1.28 GHz. The SC selective modulation formats are only set to be BPSK, 4-QAM, and 8-QAM. ONU request durations and resource demands are as follows: 80% of the requests have duration uniformly between 1t and 3t, the remaining are uniformly from 10t to 15t. 40% of ONUs is evenly distributed at [0.08 Gb/s, 0.16 Gb/s], 40% is evenly distributed at [0.32 Gb/s, 0.64 Gb/s], 20% is evenly distributed at [1.28 Gb/s, 2.56 Gb/s]. The "image" used by the DRL agent is 20t long and each experiment lasts 50t. The agent allocates for a subset of $M$ ONUs (we use $M = 8$), while observe other ONU requests in backlog queue. We set the length of the backlog to be 64 ONU requests. At each training iteration, we use 50 different request sets and run 10 Monte Carlo simulations in parallel for each request set. The DRL-based scheme with flexible modulation format is compared against two benchmarks: 1) Random, which selects requests randomly, 2) SRF (Short Request First), which servers ONU requests in ascending order of their duration. Note that the two benchmarks use the fixed 4-QAM format for each SC, which is the minimal modulation format to meet the maximum bandwidth demand 2.56 Gb/s. The weights $\alpha$ and $\beta$ for objective and reward are all set to be 1.

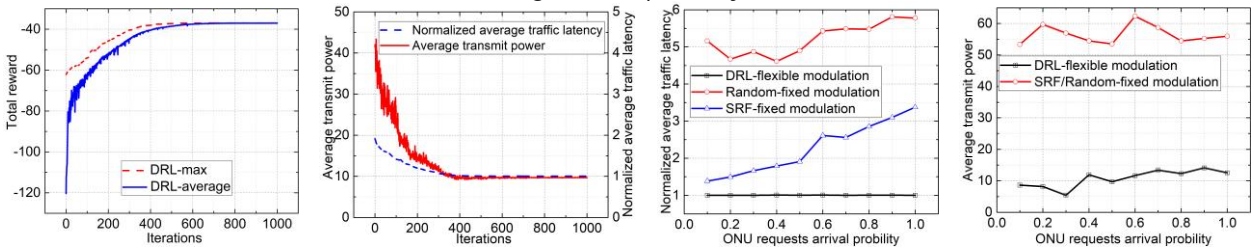


Fig. 4. The convergence property of DRL-based algorithm (a) total reward, (b) average transmit power and normalized average slowdown; DRL with flexible modulation vs. benchmarks with fixed modulation (c) normalized average slowdown, (d) average transmit power.

Fig. 4 (a) presents the trend of the maximum reward and the average reward at λ=0.5. As the iteration progresses, both values increase with the continuous training of DNN, and the gap between two values gradually converges a stable value, which indicates the system has reached an optimal state in this moment. As shown in Fig.4 (b), higher rewards, which is what DRL-based algorithm explicitly optimizes for, directly brings out the improvements in average traffic delay and transmit power of ONUs. The simulation results in Fig. 4 (c) and (d) show that the DRL-based scheme achieve the near-optimal traffic delay, while significantly decrease the transmission power by above 80% for all request arrival rate λ.

## 5. Conclusion

This work presented a DRL-based policy for power efficient DSA in OFDMA-PONs to reduce traffic delay and transmit power for the ONU requests. The simulation results proved that the proposed algorithm can reach the near-optimal average traffic delay and significantly improve the ONU transmit power for all request arrival rate λ.

## 6. References


[1] X. Gong, et al., "Joint resource allocation and software- based reconfiguration for energy- efficient OFDMA-PONs,"JOCN **10**, 75-85 (2018).
[2] W. You, et al., "Power efficient dynamic bandwidth allocation algorithm in OFDMA-PONs," JOCN **5**, 1353-1360 (2013).
[3] H. Mao, et al., "Resource Management with Deep Reinforcement Learning ,"ACM HotNets, 2016.
[4] M. R. Raza, et al., "A Slice Admission Policy Based on Reinforcement Learning for a 5G Flexible RAN," ECOC,2018, pp. 1-3.
[5] Z. Gao, et al., "Deep Reinforcement Learning for BBU Placement and Routing in C-RAN," OFC,2019, pp. 1-3.