2022 Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC)

A Deep Reinforcement Learning Policy for Joint Antenna Selection and Radio Resource Block Allocation in a TWDM-PON Based Front-haul with Massive MIMO

Yunwu Wang
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: wangyunw@seu.edu.cn

Min Zhu*
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: minzhu@seu.edu.cn

Jiahua Gu
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: gujiahua@seu.edu.cn

Xiang Liu
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: xianglliu@seu.edu.cn

Weidong Tong
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: weidongtong@seu.edu.cn

Jiao Zhang
National Mobile Communications Research Laboratory
Southeast University
Purple Mountain Laboratories
Nanjing, China
e-mail: jiaozhang@seu.edu.cn

2022 Asia Communications and Photonics Conference (ACP) | 978-1-6654-8155-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACP55869.2022.10088731

***Abstract*—We propose a deep reinforcement learning-based policy for massive multiple input multiple output (MIMO) enabled beamforming in a front-haul network. The simulation results show the proposed algorithm can achieve better performance than heuristics.**

Keywords-Massive MIMO; beamforming; TWDM-PON based front-haul; beam antenna array mapping; deep reinforcement learning

I. INTRODUCTION

Massive multiple input multiple output (mMIMO) technology, using beamforming and multiplexing in spatial domain, can improve spectral efficiency, which is considered to be a promising technique for the next generation of wireless communications [1]. It allows multiple antennas to transmit the same data to serve a user to improve the signal strength and quality. However, beamforming may incur tremendous redundant data over front-haul in next generation radio access networks (NG-RANs), which poses a great challenge on network operators to make an optimal optical (e.g., wavelength) and radio resources (e.g., antenna and radio resource block (RB)) allocation [2].

Conventional joint optical and radio resources optimization is based on integer linear programming (ILP) or heuristic algorithms, where the ILP is usually built upon the mathematical principles of optical-wireless access networks. However, the required running time of the ILP algorithm prohibits real-time resource allocation in practical networks. In addition, heuristic algorithms are difficult to achieve optimal performance among optical and radio resources in NG-RANs. This is because heuristic-based algorithms that employ predefined procedures prone to stop searching for optimal algorithms once they get an available solution [3]. Recently, the joint optimization problem of optical and radio resources is expected to be enlightened by machine learning-based algorithms, especially deep reinforcement learning (DRL) [4]. Moreover, it has a self-learning ability and can make an optimal resource allocation for the current network environment. Thus, in this paper, we adopt the DRL policy for solving the optical and radio resources allocation issue.

Very recently, the authors in [5] investigated the tradeoff between the front-haul bandwidth and radio RB utilization for mMIMO enabled beamforming in a time and wavelength division multiplexed and passive optical network (TWDM-PON). However, for the mapping of beam antenna array (BAA) request, the antenna is considered simply as a resource pool, and the specific physical structure of the antenna array is not considered from a realistic view. Thus, the antenna selection does not take into account the constraints of antenna array resource allocation, which is usually unreasonable. This motivates us to investigate the

antenna selection as a 2D space for the BAA mapping model in the mMIMO system.

In this paper, we propose a DRL policy, and apply the DRL policy for the joint optimization problem of wavelength, 2D antenna, and radio RB resource for mMIMO beamforming in a TWDM-PON based NG-RANs. The main contributions of this paper are listed as follows. 1) We propose a BAA mapping model and transform it into an evolved 3D bin-packing problem to solve the joint optimization problem between wavelength, 2D antenna and radio RB resource. 2) We design a DRL policy according to the different antenna selection policies for the 3D BAA mapping model to optimize front-haul bandwidth and the radio RB utilization. 3) The extensive simulation results demonstrate that the proposed DRL policy can effectively optimize the front-haul bandwidth and radio RB utilization.

II. Architecture and Model Description

In this section, we will briefly describe the system model, the 3D BAA mapping model for beamforming and present some preliminaries regarding the proposed policy.

A. TWDM-PON Based System Model

As shown in Fig. 1, we consider a TWDM-PON based front-haul for mMIMO system. In the base station (BS), a large-scale antenna array can be divided into several sub-antenna arrays (sub-AAs), and these sub-AAs are linked to the different remote radio unit (RRU), respectively. For a large-scale antenna array, these RRUs are associated with an optical network unit (ONU) via a time division multiplexed (TDM) switch. These RRUs are deployed near the ONU side. In the TWDM-PON based front-haul, an ONU can be assigned one or multiple wavelengths, and a wavelength can be shared among multiple ONUs. The ONUs are responsible to receive downstream signals using a specified optical wavelength resource and to transmit upstream signals from multiple antennas. The enhanced common public radio interface (eCPRI) signals are further processed in the distributed unit (DU) pool after passing through the front-haul network. The DU pool is usually placed together with an optical line terminal (OLT), whose main function is to receive upstream data from the DU and send downstream data to one or multiple ONUs by broadcasting.

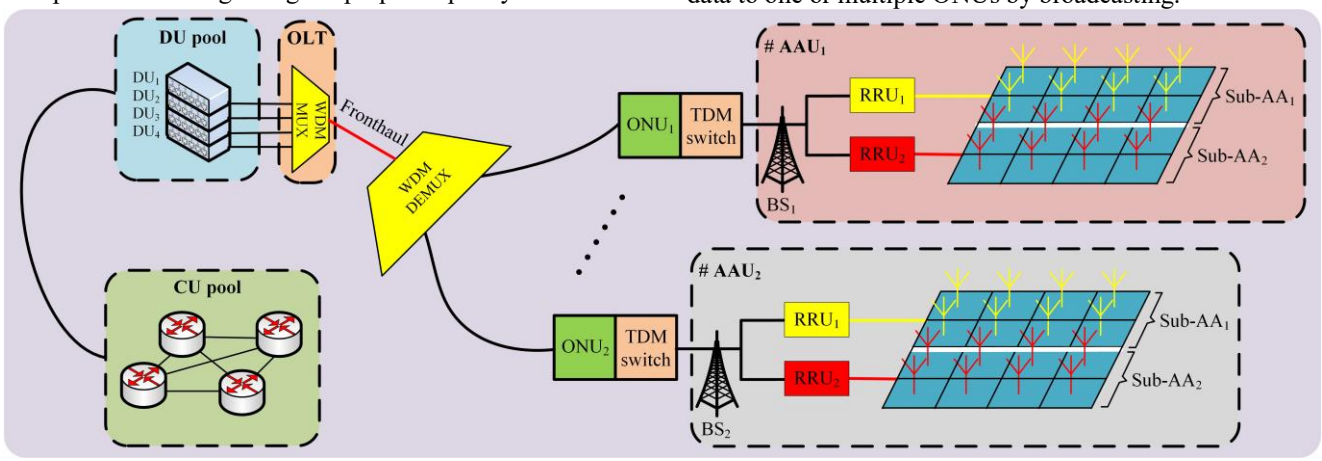

Figure 1. A TWDM-PON based front-haul for mMIMO system.

From a realistic view, a BAA request is mapped into a selected 2D antenna array in a large-scale antenna array. When the selected antenna array is just assigned within a sub-AA, which is connected to an RRU, it would not generate the signal duplication being transmitted over the front-haul. On the contrary, when the 2D antenna array for a BAA request may come from multiple RRUs, and since the RRUs are independent of each other, it is necessary for all RRUs involved in the BAA to transmit the BAA signal, resulting in redundant data transmission over the front-haul. In addition, considering the allocation of radio RB resources, adjacent BAAs could be allocated non-overlapped radio RBs on the time/frequency domains to reduce the interference between BAAs. In contrast, non-adjacent BAAs could use the same radio RBs to improve the radio RB utilization.

B. Functional Split RAN Architecture for Beamforming

To decrease front-haul bandwidth and latency in the NG-RAN, some radio signals processing functions of the low physical layer (Low-PHY) are transferred to the RRU site. For the functional spilted RAN architecture, the higher layer functions such as packet data convergence protocol (PDCP) and radio link control (RLC) are processed at the centralized unit (CU), and media access control (MAC) functions and forward error correction (FEC) encoding are processed at the DU. The CU is connected to DU through the F1 interface defined by 3GPP [6]. This paper focuses on the splitting point between the DU and the active antenna unit (AAU), and the front-haul bandwidth in the TWDM-PON can be calculated by [7]

$$F_{I_D} = 1/T_{TTI} \times N_{mcs} \times N_{sys} \times N_{sc} \times N_{rb} \times N_{mimo} \quad (1)$$

where T_{TTI} denotes a transmission time interval (TTI), N_{mcs} denotes the modulation and coding (MCS), N_{sys} denotes the number of orthogonal frequency division multiplexing (OFDM) symbols in a TTI, N_{sc} is the number of sub-carriers per RB, N_{rb} is the number of occupied RBs, and N_{mimo} is the number of MIMO streams. For each BAA request, N_{mimo} equals 1.

C. Joint Antenna Selection and Radio RB Allocation

In this paper, we formulate a BAA mapping model to solve a joint optimization problem for the antenna selection and radio RB allocation, and transform it into an evolved 3D bin-packing problem. For each BAA request, as shown in Fig. 2, the 3D size of the bin is the size in X- and Y-axis of the selected 2D antenna array and the number of the allocated radio RB in Z-axis, respectively. To intuitively demonstrate the 3D BAA mapping for mMIMO enabled beamforming, an example is presented in Fig. 2. For simplicity and clarity, only two RRUs and three BAA requests are drawn. As shown in Fig. 2, the BAA_1 request is denoted as (ax, ay, rb) = (3, 4, 4), where the ax, ay, rb and denote the value in X- and Y-axis of the 2D antennas array to be allocated for the BAA_1 request and the required radio RBs in Z-axis, respectively. Similarly, the requests BAA_2 and BAA_3 are represented as (3, 4, 4) and (4, 3, 4), respectively. In Fig. 2, considering that the BAA_2 and BAA_3 have overlapping coverage, the two BAAs can use all or part of the same 2D antenna array, but should be allocated different radio RBs. It helps to improve the utilization of antenna resources. In addition, since the BAA_1 and BBA_2 requests have different coverage, the two requests should use the same radio RBs to enhance the radio RB utilization.

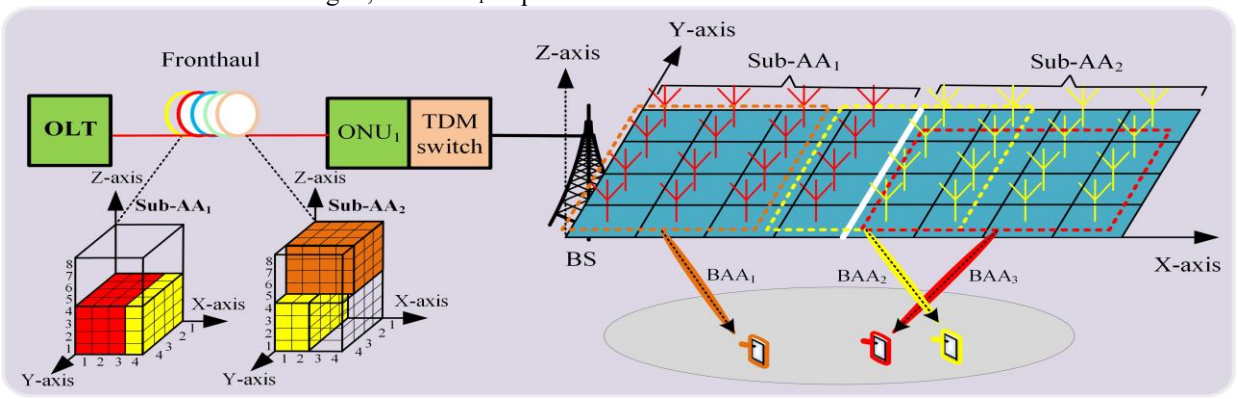

Figure 2. An example of a 3D BAA mapping model in a TWDM-PON based front-haul.

III. Description of DRL-based Policy

Joint antenna selection and radio RB allocation problems are commonly optimized by the heuristic algorithms [2], [5]. However, they are difficult to achieve optimal performance among wavelength, 2D antenna and radio RB resources in NG-RANs. Therefore, we propose a DRL-based policy to optimize the front-haul bandwidth and radio RB utilization. The structure of DRL-based policy is shown in Fig. 3. At each decision moment, the agent acquires a state from the environment. The deep neural network (DNN) maps the state inputs and the fixed requirement of BAA requests into multiple Q-values. Each Q-value is related to an action respectively. Then the action with maximum Q-value is taken. Finally, a reward is obtained from the environment after the action is performed. The detailed training process of our proposed DRL policy is shown in algorithm 1. In lines 1-2, we first sort the BAA requests in set M in a descending order of the total number of requested RBs ($ax_m \times ay_m \times rb_m$), where ax_m and ay_m denote the number of selected 2D antenna array for BAA m in X- and Y-axis, respectively, and rb_m denotes the number of radio RBs for the BAA m in Z-axis, and serve the BAA requests one by one from the top. Then, we find an available radio RB set $\{RB\text{-}RB_K\}$ that could be allocated to BAA m without causing interference with other BAAs (lines 3-5).

Algorithm 1: Procedures of DRL training

1. **for** each episode **do**
2. Sort $m \in M$ in a descending order of the total number of requested RBs ($ax_m \times ay_m \times rb_m$).
3. **for** $t = 1, T$ **do**
4. **for** $m \in M$ **do**
5. Find a subset K in M', where K's coverage overlaps with BAA m, M' denotes allocated BAA set. RB_K is a radio RB set allocated to K. $\{RB\text{-}RB_K\}$ is the available radio RB set for BAA m. And observe state s_t.
6. Simple c~Uniform(0,1)
7. **if** $c > \varepsilon$, **then**
8. Select an action $a_t \in \arg\max(s_t, a)$
9. **else**
10. Select an action $a_t \in A$ at random
11. **end if**
12. Execute action a_t, observe reward r_t and next states s_{t+1}.
13. Store transition (s_t, a_t, r_t, s_{t+1}) in D
14. Minibatch sample from D for experience (s_t, a_t, r_t, s_{t+1}), update DNN with the loss function.
15. **end for** **end for** **end for**
16. Compute the front-haul bandwidth and radio RB utilization for all RRU.

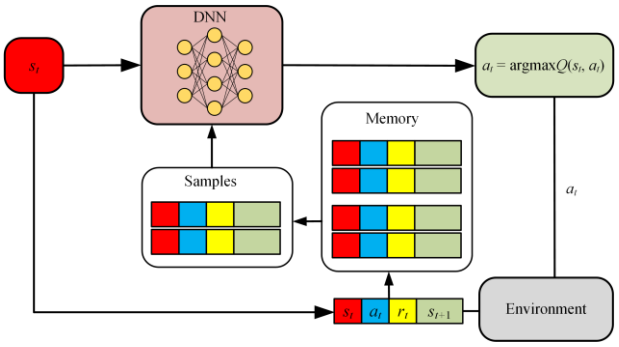

Figure 3. Structure of the proposed DRL-based policy.

State: In line 5, the agent obtains the current state. For each BAA m, the DRL generates four policies based on the

different antenna selection policies. The first policy allows the BAA m to be allocated to different RRUs and finds the lowest available antenna in X-axis. The second policy allows the BAA m to be allocated to different RRUs and finds the lowest available antenna in Y-axis. The third policy is to find the available antenna with smallest antennas index among all antennas. The forth policy does not allow the BAA m to be allocated to different RRUs and finds the lowest available antenna in X-axis. The final policy does not allow the BAA m to be allocated to different RRUs and finds the lowest available antenna in Y-axis. In our model, antennas and radio RBs allocated to a BAA m have to be continuous, which means the above four policies can be represented by their least allocated RB index and least allocated antenna index, as $\{(x_i, y_i, z_i)|_{i\in[1,5]}\}$. In addition, to further evaluate every policy, we calculate the number of used sub-AA number and used antenna number by these policies as $\{(n_i^{sub}, n_i^{a})|_{i\in[1,5]}\}$. As a result, for a given BAA m (ax_m, ay_m, rb_m), the state representation is defined as follows:

$$s_t = \{(x_i, y_i, z_i, n_i^{sub}, n_i^{a})|_{i\in[1,5]}, ax_m, ay_m, rb_m\} \quad (2)$$

Action: In lines 6-11, the agent selects an action with ε strategy. In each step, the proposed DRL fetches a BAA in M until M is empty. For each BAA m, the agent observes the state according to (2) and chooses an allocation policy from the five policies as mentioned described above.

Reward: The average cost is presented as follows:

$$AC = \alpha \cdot \left(\sum_{m\in M} \sum_{q\in Q} fh_m \times Y_q^m \right) + \beta \cdot \left(\sum_{(i,j)\in N} V_{(i,j)} \right) \quad (3)$$

where fh_m means front-haul bandwidth required by BAA m, Y_q^m and $V_{(i,j)}$ denote whether BAA m occupies sub-AA q and whether antenna (i, j) is occupied, respectively. The average cost has two parts: the first part is to minimize the front-haul bandwidth, and the second part is to maximize the radio RB utilization which is then transformed to minimize the number of occupied antennas. The two parts are linearly summed up by multiplying the weighting factors α and β. In line 12, the agent obtains the reward. For any BAA m, the average cost is calculated twice before and after scheduling BAA m as AC_m and AC_{m+1}, respectively. Hence, the reward can be produced as $r_t = AC_m - AC_{m+1}$, which reflects the negative average cost consumed by this BAA m. The more resources the BAA m uses, the large penalty the DRL agent will receive.

Training: In this work, we implement double deep Q network (DDQN) as our DRL algorithm. All inputs are discrete variables so we transform them into one-hot code to converge better. In lines 13-15, we record data (s_t, a_t, r_t, s_{t+1}) in the memory D. To train the DRL model, the parameters of DNN is updated by $loss = E\{[r_t + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]^2\}$, where $r_t + \gamma \max Q(s_{t+1}, a_{t+1})$ is the optimal Q-value, $Q(s_t, a_t)$ denotes the Q-value before update [7].

IV. Performance Evaluation

In this section, extensive simulations are performed to evaluate the proposed DRL policy. In our simulation, for the simulation setup of the wireless part, we consider that each wireless RB has a frequency range of 180 kHz, 12 subcarriers and 7 OFDM symbols are transmitted on each TTI with MCS of 7, i.e., $N_{sc} = 12$, $N_{sys} = 7$, $N_{mcs} = 7$. For the sake of simplicity, this paper considers only the situation of a single BS with one large-scale antenna array. In a large-scale network, we have 16 × 16 antenna divided into 4 × 4 sub-AAs and each sub-AA includes 4 × 4 antennas, each antenna has 1000 radio RB, For each BAA, the antenna requests are $\{(ax \times ay)|ax \in [2, 4], ay \in [2, 4]\}$ and the RB request of each BAA is random in [10, 50]. In addition, each BAA has two characteristics: angle and direction. The BAA requests with antenna request numbers [1-4]/[5-8]/[9-12]/[13-16] have a beam coverage of 20°/15°/10°/5°, respectively. In the simulation, the minimum beam angle is 5°. To accurately calculate the extent of overlap between BAAs, we consider an omnidirectional antenna system providing 360° radiation in the horizontal direction, and the coverage of the antenna array can be divided into 360 small regions with an interval of 1°.

Algorithm 2: Inter-Subarray BAA Mapping Algorithm (Inter-SA)

1. for $m \in M$ **do**
2. Find a subset A_c in A', in which every antenna has a same consecutive RB set that satisfies $RB_{AC} \in \{RB\text{-}RB_K\}$, $|RB_{AC}| \geq rb_m$, where A' denotes allocated antenna set, then
4. if $|A_{cx}| \geq ax_m$, $|A_{cy}| \geq ay_m$, **then**
5. Select a subset A_m in A_c, allocate rb_m consecutive RBs (i.e., RB_{AM}) from the available RB in A_m.
6. elif $|A_{cx}| \geq ax_m$, $|A_{cy}| < ay_m$, **then**
7. Allocate rb_m consecutive RBs (i.e., RB_{Ac}) from the lowest available RB in A_c. Allocate an empty RRU in the Y-axis in all RRU, select an antenna set A'_c, allocate rb_m consecutive RBs in A'_c as same as RB_{Ac}.
8. elif $|A_{cx}| < ax_m$, $|A_{cy}| \geq ay_m$, **then**
9. Allocate rb_m consecutive RBs (i.e., RB_{Ac}) from the lowest available RB in A_c. Allocate an empty RRU in the X-axis in all RRU, select an antenna set A'_c, allocate rb_m consecutive RBs in A'_c as same as RB_{Ac}.
10. else
11. Allocate an empty RRU, select an antenna set A_m, allocate rb_m consecutive RBs (i.e., RB_{AM}) from the lowest available RB in A_m.
12. end if end for
13. Compute the front-haul bandwidth and radio RB utilization for all RRU.

We compare our proposed DRL policy with two benchmark heuristics, including the Intra-subarray BAA mapping algorithm (Intra-SA) and the Inter-subarray BAA mapping algorithm (Inter-SA). The key idea of the Intra-SA is to find the RRU that satisfies the current BAA requests in the occupied RRUs, and if not found, a new RRU is occupied to serve the current BAA request and update the available resources of the RRUs. The Intra-SA can minimize the front-haul bandwidth. However, it requires more antennas to be employed, resulting in lower radio RB utilization. Therefore, we consider that the 2D antenna array of each BAA may come from adjacent RRUs to improve radio RB utilization as much as possible. The process of the Inter-SA is shown in algorithm 2.

Fig. 4 (a) demonstrates the results of the average cost calculated by Eq. (3). As shown in Fig. 4(a), it can be observed that the proposed DRL-based algorithm achieves the lowest average cost compared with the benchmark algorithms, which proves the effectiveness of the proposed

DRL-based algorithm. To reveal the fundamental reasons, we compare the results of front-haul bandwidth, antenna consumption and the radio RB utilization in Figs. 4(b), 4(c) and 4(d), respectively. Fig. 4(b) shows the front-haul bandwidth versus the number of BAA requests. As shown, the front-haul bandwidth requirements increase dramatically as the number of BAA requests increases. The DRL-based is higher than the Intra-SA but significantly lower than the Inter-SA. This reason is that the Intra-SA maps the 2D antennas of each BAA request to the same RRU to achieve the optimal front-haul bandwidth.

Fig. 4(c) shows the number of used antennas versus the number of BAA requests. As shown, DRL-based achieves the lowest antenna consumption compared with Intra-SA and Inter-SA. This is because the Intra-SA attributes each BAA request to the same RRU, which results in using more antennas. In the three algorithms, Intra-SA uses the most antennas since the Intra-SA is essentially an algorithm that reduces the front-haul bandwidth at the cost of using additional antennas. From Figs. 4(b) and 4(c), we confirm that to achieve the minimum average cost, the DRL agent would trade some front-haul bandwidth for much improved antenna utilization. Fig. 4(d) shows the radio RB utilization versus the number of BAA requests. As shown, the DRL-based employs a higher radio RB utilization while the Intra-SA employs the lowest radio RB utilization. Accordingly, the DRL-based achieves the simultaneous optimization of the front-haul bandwidth and radio RB utilization.

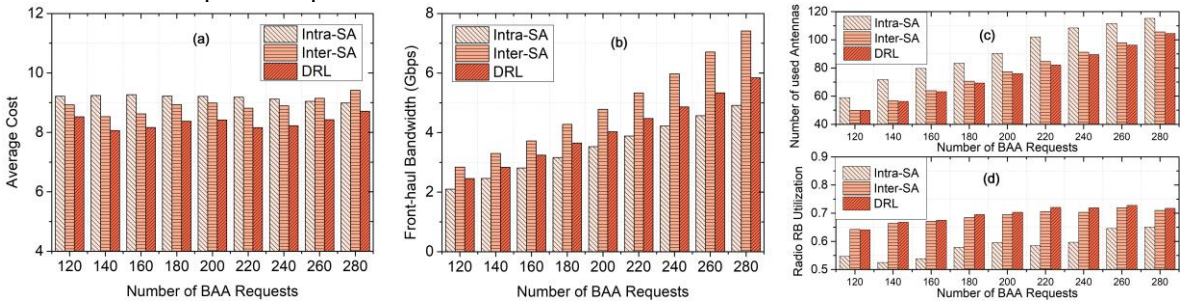

Figure 4. Number of BAA requests vs. (a) average cost, (b) Front-haul bandwidth (Gbps), (c) Number of the utilized antennas and (d) Radio RB utilization in large-scale network.

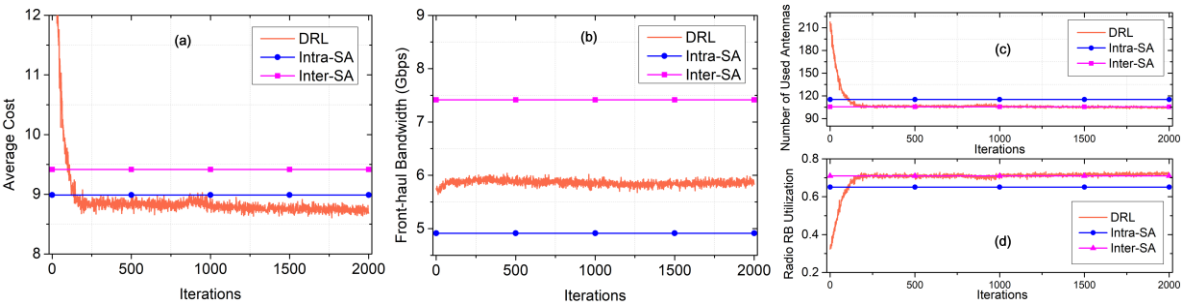

Figure 5. Training results of DRL vs. benchmark heuristics in the large-scale network: (a) average cost, (b) Front-haul bandwidth (Gbps), (c) Number of the utilized antennas and (d) Radio RB utilization.

Fig. 5 shows the training process of DRL in large-scale network with iteration-by-iteration average cost (Fig. 5(a)), front-haul bandwidth (Fig. 5(b)), antenna consumption (Fig. 5(c)) and the radio RB utilization (Fig. 5(d)), where the number of BAA requests equal 280. Fig. 5(a) shows the trend of average cost against training iterations for DRL-based algorithm. The training begins with high average cost because of the random exploration, then decreases quickly for improving algorithm, and finally converges to a cost average cost at 8.7. This means that good convergence performance of DRL-based algorithm is achieved. It is observed that by iterations 200, the average cost obtained by trained by DRL is lower than the other benchmark heuristics and the training curves in Figs. 5(a), 5(b), 5(c) and 5(d) become flatten and converged after iteration 200.

V. CONCLUSION

In this paper, a TWDM-PON based front-haul for mMIMO system was investigated. Specifically, we introduced the system architecture and the BAA mapping model for beamforming, and transformed it into an evolved 3D bin-packing problem. Then, a DRL-based policy was proposed to balance between minimizing front-haul bandwidth and maximizing radio RB utilization. The extensive simulation results demonstrated that the proposed DRL-based policy can effectively optimize the front-haul bandwidth and radio RB utilization.

REFERENCES

[1] E. G. Larsson, et al., "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[2] Y. Xiao, et al., "Integrated resource optimization with WDM-based fronthaul for multicast-service beam-forming in massive MIMO-enabled 5G networks," *Photonic Netw. Commun.*, vol. 37, no. 3, pp. 349-360. Mar. 2019.

[3] E. A. Varvarigos, et al., "Algorithmic aspects in planning fixed and flexible optical networks with emphasis on linear optimization and heuristic techniques," *J. Lightw. Technol.*, vol. 32, no. 4, pp. 681-693, Feb. 2014.

[4] S. Shen, et al., "DRL-based channel and latency aware radio resource allocation for 5G service-oriented RoF-MmWave RAN," *J. Lightw. Technol.*, vol. 39, no. 18, pp. 5706-5714, Sep. 2021.

[5] J. Zhang, et al., "Joint wavelength, antenna, and radio resource block allocation for massive MIMO enabled beamforming in a TWDM-PON based fronthaul," *J. Lightw. Technol.*, vol. 37, no. 4, pp. 1396-1407, Feb. 2019.

[6] 3GPP TS 38.470, "F1 general aspects and principles," V15.1.0, Sophia Antipolis, France, 2018. [Online]. Available: http://www.3gpp.org/DynaReport/38-series.htm.

[7] K. Miyamoto, et al., "Analysis of mobile fronthaul bandwidth and wireless transmission performance in split-PHY processing architecture," *Opt. Express*, vol. 24, no. 2, pp. 1261–1268, Jan. 2016.

[8] Z. Gao et al., "Deep reinforcement learning-based policy for baseband function placement and routing of RAN in 5G and beyond," *J. Lightw. Technol.*, vol. 40, no. 2, pp. 470-480, Jan. 2022.