Model-Free Dynamic Traffic Steering for Multi-Link Operation in IEEE 802.11be

Ching-Lun Tai[1], Mark Eisen[2], Dmitry Akhmetov[2], Dibakar Das[2], Dave Cavalcanti[2], and Raghupathy Sivakumar[1]
[1]*School of Electrical and Computer Engineering, Georgia Institute of Technology*, Atlanta, GA, United States
[2]*Intel Corporation*, Hillsboro, OR, United States
Email:{ctai32, siva}@gatech.edu, mark.eisen@ieee.org, {dmitry.akhmetov, dibakar.das, dave.cavalcanti}@intel.com

***Abstract*—The IEEE 802.11be extremely high throughput (EHT) amendment (commercially known as Wi-Fi 7) aspires to further improve the network performance on data rate, latency, and reliability. As a core capability in the IEEE 802.11be EHT amendment, multi-link operation (MLO) defines the multi-link device (MLD) architecture with the simultaneous transmit and receive (STR) functionality for both station (STA) and access point (AP) to achieve concurrent operations across multiple available links. To unleash the full potential of MLO, it is essential that the AP MLD create proper dynamic traffic steering for STA MLDs. Therefore, in this paper, we consider MLO with the STR functionality and propose two application-agnostic model-free dynamic traffic steering methods. Specifically, we develop an adaptive scoring heuristic algorithm based on a normalized weighted score featuring lightweight computational complexity and a deep reinforcement learning (DRL) approach based on standard soft actor-critic (SAC) featuring robust adaptability. Simulation results demonstrate the effectiveness of the proposed model-free dynamic traffic steering methods in terms of reward convergence and average network throughput. Besides, we investigate the load balancing ability of the proposed model-free dynamic traffic steering methods.**

***Index Terms*—Multi-link operation (MLO), dynamic traffic steering, adaptive scoring, deep reinforcement learning (DRL), soft actor-critic (SAC)**

I. Introduction

Over the past two decades, Wi-Fi has been a widely adopted wireless local area network (WLAN) technology, providing ubiquitous connectivity around the globe. As new applications with more stringent requirements increasingly emerge in highly dense and congested wireless networks, the IEEE 802.11be extremely high throughput (EHT) amendment, which will be commercialized as Wi-Fi 7, proposes several advanced features to support high data rate, low latency, and high reliability for various scenarios [1].

Among the advanced features proposed in the IEEE 802.11be EHT amendment, multi-link operation (MLO) [2] defines a new architecture called multi-link device (MLD) with the simultaneous transmit and receive (STR) functionality toward concurrent operations across multiple available links for both station (STA) and access point (AP). Specifically, an STA MLD or an AP MLD possesses multiple interfaces, each of which manages an available link, to enable MLO. For an exploration of MLO, some previous works study its performance in terms of latency, reliability, and throughput (e.g., [3]–[5]), while some others investigate the coexistence of MLDs and legacy devices (e.g., [6]–[8]).

In order to take full advantage of MLO, the AP MLD needs to create proper dynamic traffic steering by dynamically determining the traffic steering portion of each available link, which is the portion of unallocated packets to be steered to each available link, for every STA MLD according to given information. Making this determination is challenging in practice, due to the complexities of throughput models in modern Wi-Fi networks and lack of knowledge of various parameters that affect traffic steering across available links, such as congestion levels and traffic patterns. It is thus critical to design dynamic traffic steering methods that are *model-free*, meaning that they can adapt to changes in the environment without explicit knowledge or modeling.

There are several previous works on dynamic traffic steering for MLO with regard to downlink (DL) traffic. The authors of [9] propose a multi-link congestion-aware load balancing (MCAB) dynamic traffic steering policy where the traffic steering portion of each available link is proportional to the remaining channel airtime. In [10], a deep reinforcement learning (DRL) dynamic traffic steering strategy based on modified soft actor-critic (SAC) [11] is proposed, selecting the traffic steering portion of available links from a predetermined finite discrete set for traffic from specific applications. However, the dynamic traffic steering techniques proposed in the above previous works suffer from a constrained traffic steering decision with limited possibilities and do not fully consider the effect of dynamic parameters in a realistic Wi-Fi network.

In this paper, we consider MLO with the STR functionality and propose two application-agnostic model-free dynamic traffic steering methods, which determine a flexible traffic steering portion according to multiple crucial factors in a computation-efficient manner (compared to model-based methods) that is favorable to an AP MLD with limited computational resources. Particularly, we develop an adaptive scoring heuristic algorithm based on a normalized weighted score with lightweight computational complexity and a DRL approach based on standard SAC [12] with robust adaptability. With sufficient flexibility, the proposed model-free dynamic traffic steering methods organically adapt to arbitrary MLO environments across various network/traffic configurations.

The remainder of this paper is organized as follows. Sec.

II describes the system model and problem formulation. In Sec. III, we introduce the proposed adaptive scoring heuristic algorithm and analyze its computational complexity. In Sec. IV, we deal with the problem from the DRL aspect and introduce the proposed SAC-based DRL approach. Simulation results and discussions are included in Sec. V. Finally, Sec. VI concludes the paper.

II. System Model and Problem Formulation

In this section, we present the system model and dynamic traffic steering problem for MLO in Wi-Fi.

Consider an MLO-enabled Wi-Fi network with the STR functionality that consists of an AP MLD and M STA MLDs with uplink (UL) traffic. Each MLD is equipped with L interfaces, which manage L available links. In a centralized fashion, the AP MLD needs to employ its collective information to create dynamic traffic steering coordinately across STA MLDs, which are informed and scheduled by trigger frames sent from the AP MLD, to steer unallocated UL packets to available links. For each STA MLD, the UL packets that have been allocated to its lth available link will be transmitted from its lth interface to the lth interface of the AP MLD. An illustration of the MLO-enabled Wi-Fi network described above is shown in Fig. 1.

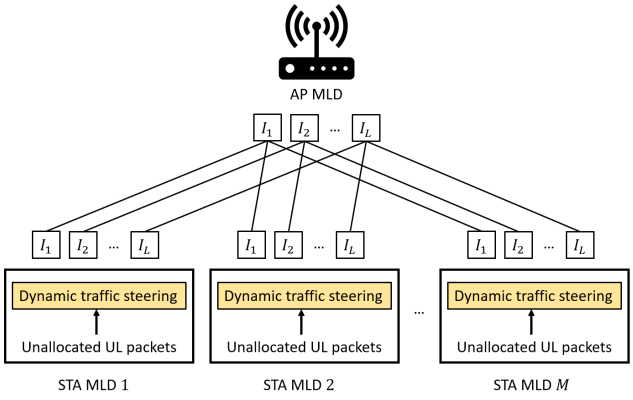


Fig. 1. An illustration of MLO-enabled Wi-Fi network, where I_l represents the lth interface and the solid lines between interfaces represent available links

To create proper dynamic traffic steering for every STA MLD that maximizes the network throughput, the AP MLD leverages multiple crucial factors such as signal-to-noise ratio (SNR), number of allocated UL packets, and channel busy time. For an available link, SNR reveals the channel quality, number of allocated UL packets implies the load level, and channel busy time indicates the channel occupancy.

Suppose there are T UL transmission windows where the STA MLDs send allocated UL packets to the AP MLD over available links. For the lth available link in the mth STA MLD, we use $\xi_{m,l}[t]$, $p_{m,l}[t]$, $c_{m,l}[t]$, and $b_{m,l}[t]$ to denote the SNR, the number of allocated UL packets, the channel busy time (measured from the previous UL transmission window), and the number of UL packets to be allocated, respectively, at the start of the tth UL transmission window, and $x_{m,l}[t]$ and $d_{m,l}[t]$ the number of allocated UL packets that have been transmitted and discarded, respectively, during the tth UL transmission window, $l = 1, 2, ..., L, m = 1, 2, ..., M, t = 1, 2, ..., T$. Then, $p_{m,l}[t]$ can be expressed by the following evolution equation:

$$p_{m,l}[t] = \max\{p_{m,l}[t-1] + b_{m,l}[t-1] - x_{m,l}[t-1] - d_{m,l}[t-1], 0\}. \quad (1)$$

It should be noted that both the number of allocated UL packets that have been transmitted $x_{m,l}$ and the number of UL packets to be allocated $b_{m,l}$ depend on the SNR $\xi_{m,l}$, the number of allocated UL packets $p_{m,l}$, and the channel busy time $c_{m,l}$, i.e.,

$$x_{m,l} = x_{m,l}(\xi_{m,l}, p_{m,l}, c_{m,l}), \quad (2)$$

$$b_{m,l} = b_{m,l}(\xi_{m,l}, p_{m,l}, c_{m,l}). \quad (3)$$

At the start of each UL transmission window, there will be unallocated UL packets in every STA MLD that need to be steered to available links. According to the information $(\xi_{m,l}, p_{m,l}, c_{m,l})$, $l = 1, 2, ..., L, m = 1, 2, ..., M$, the AP MLD needs to determine the traffic steering portion $\alpha_{m,l} \in [0, 1]$, which is the portion of unallocated UL packets to be steered to the lth available link in the mth STA MLD. Denote the network throughput during the tth UL transmission window as δ_t, $t = 1, 2, ..., T$.

Based upon the above system model, we formulate the following problem to be tackled in this work: Suppose there are T UL transmission windows. At the start of the tth UL transmission window, given SNR $\xi_{m,l}[t]$, number of allocated UL packets $p_{m,l}[t]$, and channel busy time $c_{m,l}[t]$, determine the traffic steering portion of the L available links for unallocated UL packets in the mth STA MLD, $(\alpha_{m,1}, \alpha_{m,2}, ..., \alpha_{m,L}) \in [0, 1]^L$, where $\sum_{l=1}^{L} \alpha_{m,l} = 1$, with the objective of maximizing the average network throughput $\frac{1}{T}\sum_{t=1}^{T} \delta_t$ over the T UL transmission windows.

III. Adaptive Scoring Heuristic Algorithm

In this section, we propose an adaptive scoring heuristic algorithm to the MLO dynamic traffic steering problem formulated in Sec. II and analyze its computational complexity.

A. Algorithm Overview

For every STA MLD, the AP MLD needs to examine each available link according to given information in order to determine its MLO dynamic traffic steering, i.e., the traffic steering portion of each available link, at the start of every UL transmission window. Therefore, we develop an adaptive scoring heuristic algorithm which assigns the traffic steering portion to each available link based on a normalized weighted score with the weight adapted in an iterative manner, as illustrated in Algorithm 1.

Denote the weight for the lth available link in the mth STA MLD as $w_{m,l}$. To begin with, we initialize every weight as unity, i.e., $w_{m,l} = 1, l = 1, 2, ..., L, m = 1, 2, ..., M$.

For the mth STA MLD at the start of the tth UL transmission window, the AP MLD has the corresponding information of SNR $\xi_{m,l}[t]$, number of allocated UL packets $p_{m,l}[t]$, and channel busy time $c_{m,l}[t]$ per available link, $l = 1, 2, ..., L$.

Algorithm 1: Adaptive Score-Based Traffic Steering

Initialization: $w_{m,l} = 1, l = 1, 2, ..., L, m = 1, 2, ..., M,\ \delta' = 0$
for $t = 1 : T$
for $m = 1 : M$
  Input: $\xi_{m,l}[t]$, $p_{m,l}[t]$, $c_{m,l}[t]$, $l = 1, 2, ..., L$
  for $l = 1 : L$
    $\eta_{m,l}[t] = 1/\xi_{m,l}[t]$
  end for
  $(\eta_m, p_m, c_m) = \sum_{l=1}^{L} (\eta_{m,l}[t], p_{m,l}[t], c_{m,l}[t])$
  for $l = 1 : L$
    $(\bar{\eta}_{m,l}, \bar{p}_{m,l}, \bar{c}_{m,l}) = (\eta_{m,l}[t]/\eta_m, p_{m,l}[t]/p_m, c_{m,l}[t]/c_m)$
    $\psi_{m,l} = w_{m,l}/(1 + \bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l})$
  end for
  $\psi_m = \sum_{l=1}^{L} \psi_{m,l}$
  Output: $(\alpha_{m,1}, \alpha_{m,2}, ..., \alpha_{m,L}) = (\psi_{m,1}, \psi_{m,2}, ..., \psi_{m,L})/\psi_m$
  $\Omega_m = \{l' : \alpha_{m,l'} \geq 1/L\}$
  Obtain: δ_t
  if $t > 1$
    for l **in** Ω_m
      $w'_{m,l} = w_{m,l}$
      $w_{m,l} = \min\{\max\{w'_{m,l} \cdot \delta_t/\delta', w_{\min}\}, w_{\max}\}$
    end for
  end if
  $\delta' = \delta_t$
end for
end for

Define the reciprocal of SNR as link poorness, which can be expressed as

$$\eta_{m,l}[t] = 1/\xi_{m,l}[t]. \tag{4}$$

To fairly accommodate the effect of three factors, link poorness $\eta_{m,l}[t]$, number of allocated UL packets $p_{m,l}[t]$, and channel busy time $c_{m,l}[t]$, of different scales, we normalize them into positive values within the same range between zero and unity. Specifically, we compute the sum for each factor across the L available links as

$$(\eta_m, p_m, c_m) = \sum_{l=1}^{L} (\eta_{m,l}[t], p_{m,l}[t], c_{m,l}[t]) \tag{5}$$

and normalize the three factors for the lth available link as

$$(\bar{\eta}_{m,l}, \bar{p}_{m,l}, \bar{c}_{m,l}) = (\eta_{m,l}[t]/\eta_m, p_{m,l}[t]/p_m, c_{m,l}[t]/c_m). \tag{6}$$

For each of the three normalized factors $\bar{\eta}_{m,l}, \bar{p}_{m,l}, \bar{c}_{m,l} \in [0, 1]$, a larger value implies a more negative effect on the lth available link. Accordingly, we use the product of the three normalized factors, $\bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l} \in [0, 1]$, to represent their joint effect on the lth available link. Note that the joint effect $\bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l}$ is null when any constitutive normalized factor (which imposes a negative effect) is equal to a perfect zero.

For an assessment of the lth available link, we compute its corresponding weighted score as

$$\psi_{m,l} = w_{m,l}/(1 + \bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l}), \tag{7}$$

which is the weight $w_{m,l}$ over the discount term $1 + \bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l}$. It should be noted that the weighted score $\psi_{m,l}$ is equal to the weight $w_{m,l}$ when the product $\bar{\eta}_{m,l} \cdot \bar{p}_{m,l} \cdot \bar{c}_{m,l}$ is equal to zero, i.e., when the joint effect is null. Then, we normalize the weighted score for an assignment of the traffic steering portion. Namely, the traffic steering portion of the L available links is computed as

$$(\alpha_{m,1}, \alpha_{m,2}, ..., \alpha_{m,L}) = (\psi_{m,1}, \psi_{m,2}, ..., \psi_{m,L})/\psi_m, \tag{8}$$

where $\psi_m = \sum_{l=1}^{L} \psi_{m,l}$ is the sum of weighted scores across the L available links.

Define a core link as an available link whose traffic steering portion is greater than or equal to $1/L$ (the average traffic steering portion across the L available links). Then, we denote the set of indices of the core links as

$$\Omega_m = \{l' : \alpha_{m,l'} \geq 1/L\}. \tag{9}$$

At the end of the tth UL transmission window, we obtain the network throughput δ_t. Since the core links play a more significant role (compared to those which are not core links) during the UL transmission window, we adapt their weight, for a reflection of their performance, as

$$w_{m,l}^{(t+1)} = \min\{\max\{w_{m,l}^{(t)} \cdot \delta_t/\delta_{t-1}, w_{\min}\}, w_{\max}\} \in [w_{\min}, w_{\max}] \tag{10}$$

for all $l \in \Omega_m$, where $w_{m,l}^{(t)}$ and $w_{m,l}^{(t+1)}$ are the values of the weight $w_{m,l}$ for the tth and $(t+1)$th UL transmission windows, respectively, and $w_{\min}$ and $w_{\max}$ are the minimum and maximum values of the weight, respectively. Intuitively, the weight of the core links increases (or decreases) when they lead to an increase (or a decrease) in the network throughput between the current and previous UL transmission windows.

B. Computational Complexity

Following the technical overview of the proposed adaptive scoring heuristic algorithm based on a normalized weighted score (Algorithm 1), we analyze its computational complexity for the MLO dynamic traffic steering of each STA MLD at the start of a UL transmission window in terms of the number of multiplications/divisions involved.

For each of the L available links, the calculation of link poorness involves one division, the normalization of the three factors (link poorness, number of allocated UL packets, and channel busy time) involves three divisions, the computation of a weighted score involves two multiplications and one division, and the assignment of traffic steering portion (normalization of weighted score) involves one division. For each of the $\mathcal{O}(L)$ core links, the weight adaptation involves one multiplication and one division.

Therefore, the total computational complexity of Algorithm 1 for the MLO dynamic traffic steering of each STA MLD at the start of a UL transmission window is $\mathcal{O}(L)$. With its lightweight computational complexity, Algorithm 1 is a feasible solution to be deployed in the AP MLD.

IV. Deep Reinforcement Learning (DRL) Approach

In this section, we elaborate on how we deal with the problem formulated in Sec. II from the DRL perspective and propose an SAC-based DRL approach.

A. DRL on Dynamic Traffic Steering for MLO

While the proposed adaptive scoring heuristic algorithm based on a normalized weighted score (Algorithm 1) features lightweight computational complexity, its traffic steering portion assignment depends on a simplified abstraction (in terms of an adaptive weight and the discount term), which may not fully delineate the underlying intricacies in MLO.

Considering the implicit complexities in MLO, we adopt DRL, where the deployed agent learns to take an action according to the observation under a trained policy by interacting with the environment and receiving a reward. Specifically, we deploy a DRL agent in the AP MLD, and the DRL agent learns from the received reward after taking an action as a response to the observation in an MLO-enabled Wi-Fi network.

In consequence, we convert the MLO dynamic traffic steering problem formulated in Sec. II into a DRL problem defined by the following components, including observation, action, and reward. For this DRL problem, we set an episode as a collection of the T UL transmission windows, where each UL transmission window is set as a step.

1) *Observation:* While a state of the environment can be fully described by its constitutive properties, an observation perceived by a DRL agent during each step consists of only a subset of these properties. During the tth step, the DRL agent in the AP MLD perceives an observation $s_t = (s_1^{(t)}, s_2^{(t)}, ..., s_M^{(t)})$, where $s_m^{(t)} = (\xi_{m,1}[t], \xi_{m,2}[t], ..., \xi_{m,L}[t], p_{m,1}[t], p_{m,2}[t], ..., p_{m,L}[t], c_{m,1}[t], c_{m,2}[t], ..., c_{m,L}[t])$ corresponds to the SNR, the number of allocated UL packets, and the channel busy time associated with the L available links in the mth STA MLD at the start of the tth UL transmission window.
2) *Action:* Based on its observation, a DRL agent takes an action during each step to interact with the environment. During the tth step, the DRL agent in the AP MLD takes an action $a_t = (a_1^{(t)}, a_2^{(t)}, ..., a_M^{(t)})$, where $a_m^{(t)} = (\alpha_{m,1}, \alpha_{m,2}, ..., \alpha_{m,L})^{(t)}$ is the traffic steering portion of the L available links in the mth STA MLD at the start of the tth UL transmission window.
3) *Reward:* For this DRL problem, the DRL agent receives a (dense) reward from the environment during each step after taking an action and perceives the next observation. During the tth step, we set the reward received by the DRL agent in the AP MLD as $r_t = (\delta_t - \mu_t)/\mu_t$, where μ_t is the benchmark network throughput of the tth UL transmission window when unallocated packets are evenly steered to each available link (whose traffic steering portion is exactly $1/L$). Accordingly, the DRL agent in the AP MLD receives a positive (or negative) reward r_t when the achieved network throughput δ_t is greater than (or less than) the benchmark throughput μ_t during the tth step and perceives the next observation $s_{t+1} = (s_1^{(t+1)}, s_2^{(t+1)}, ..., s_M^{(t+1)})$.

With the observation, action, and reward defined above, we construct a DRL problem of dynamic traffic steering for MLO to be addressed.

B. SAC-Based DRL Approach

For the DRL agent in the AP MLD, we adopt standard SAC [12], which is a state-of-the-art DRL approach that has been successfully applied to various DRL tasks, developing an SAC-based DRL approach for the MLO dynamic traffic steering, as illustrated in Algorithm 2.

Algorithm 2: DRL SAC-Based Traffic Steering

Initialization: θ, ϕ_1, ϕ_2, $\hat{\phi}_1 = \phi_1$, $\hat{\phi}_2 = \phi_2$, $\mathcal{D} = \varnothing$
for $t = 1 : T$
  Observation: $s_t = (s_1^{(t)}, s_2^{(t)}, ..., s_M^{(t)})$
  Action: $a_t = (a_1^{(t)}, a_2^{(t)}, ..., a_M^{(t)}) \sim \pi_\theta(\cdot|s_t)$
  Reward: $r_t = (\delta_t - \mu_t)/\mu_t$
  Next observation: $s_{t+1} = (s_1^{(t+1)}, s_2^{(t+1)}, ..., s_M^{(t+1)})$
  Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$.
  if $t > U$
    Randomly sample $\{s_j, a_j, r_j, s'_j\}_{j=1}^{H}$ from $\mathcal{D}$.
    for $j = 1 : H$
      $y_j = r_j + \gamma[\min_{k=1,2} Q_{\hat{\phi}_k}(s'_j, \tilde{a}') - \beta \cdot \log \pi_\theta(\tilde{a}'|s'_j)]$,
      $\tilde{a}' \sim \pi_\theta(\cdot|s'_j)$
    end for
    for $k = 1 : 2$
      Update ϕ_k with GD by minimizing
      $F_{\phi_k} = \frac{1}{H} \sum_{j=1}^{H} [Q_{\phi_k}(s_j, a_j) - y_j]^2$.
    end for
    Update θ with GD by minimizing
    $F_\theta = \frac{1}{H} \sum_{j=1}^{H} [\min_{k=1,2} Q_{\phi_k}(s_j, \tilde{a}_\theta(s_j)) -$
    $\beta \cdot \log \pi_\theta(\tilde{a}_\theta(s_j)|s_j)]$, $\tilde{a}_\theta(s_j) \sim \pi_\theta(\cdot|s_j)$.
    for $k = 1 : 2$
      $\hat{\phi}'_k = \hat{\phi}_k$; $\hat{\phi}_k = \rho \cdot \hat{\phi}'_k + (1 - \rho)\phi_k$
    end for
  end if
end for

In the beginning, we initialize a policy π_θ with parameter θ (actor), two Q-networks Q_{ϕ_1} and Q_{ϕ_2} with parameters ϕ_1 and ϕ_2 (critics), two target networks $Q_{\hat{\phi}_1}$ and $Q_{\hat{\phi}_2}$ with parameters $\hat{\phi}_1$ and $\hat{\phi}_2$, and an empty replay buffer $\mathcal{D}$.

During the tth step, the DRL agent in the AP MLD perceives an observation s_t, takes an action $a_t \sim \pi_\theta(\cdot|s_t)$, receives a reward r_t, and perceives the next observation s_{t+1} for the next step. Then, an experience tuple (s_t, a_t, r_t, s_{t+1}) is stored into the replay buffer $\mathcal{D}$.

After the first U steps, the parameters θ, ϕ_1, ϕ_2, $\hat{\phi}_1$, and $\hat{\phi}_2$ maintained by the DRL agent in the AP MLD will be updated for each step. During each update, we randomly sample a mini-batch of H experience tuples $\{(s_j, a_j, r_j, s'_j)\}_{j=1}^{H}$ from the

replay buffer $\mathcal{D}$. With the jth experience tuple (s_j, a_j, r_j, s'_j), we compute the jth target as

$$y_j = r_j + \gamma[\min_{k=1,2} Q_{\hat{\phi}_k}(s'_j, \tilde{a}') - \beta \cdot \log \pi_\theta(\tilde{a}'|s'_j)],\ \tilde{a}' \sim \pi_\theta(\cdot|s'_j), \quad (11)$$

where $\gamma \in [0,1]$ is the discount factor and $\beta \in [0,1]$ is the automatically tuned entropy regularization coefficient for a control of the explore-exploit tradeoff. For the two Q-networks Q_{ϕ_1} and Q_{ϕ_2} (critics), we update their parameters ϕ_1 and ϕ_2 with gradient descent (GD) by minimizing the loss function

$$F_{\phi_k} = \frac{1}{H}\sum_{j=1}^{H}[Q_{\phi_k}(s_j, a_j) - y_j]^2, \text{ for } k = 1, 2. \quad (12)$$

For the policy π_θ (actor), we update its parameter θ with GD by minimizing the loss function

$$F_\theta = \frac{1}{H}\sum_{j=1}^{H}[\min_{k=1,2} Q_{\phi_k}(s_j, \tilde{a}_\theta(s_j)) - \beta \cdot \log \pi_\theta(\tilde{a}_\theta(s_j)|s_j)],$$
$$\tilde{a}_\theta(s_j) \sim \pi_\theta(\cdot|s_j). \quad (13)$$

Finally, for the two target networks $Q_{\hat{\phi}_1}$ and $Q_{\hat{\phi}_2}$, we update their parameters $\hat{\phi}_1$ and $\hat{\phi}_2$ by computing

$$\hat{\phi}_k^{\text{upd}} = \rho \cdot \hat{\phi}_k + (1-\rho)\phi_k, \text{ for } k = 1, 2, \quad (14)$$

where $\rho \in [0,1]$ is the interpolation factor and $\hat{\phi}_k^{\text{upd}}$ is the updated value of parameter $\hat{\phi}_k$.

With its robust adaptability, the proposed SAC-based DRL approach (Algorithm 2) helps the DRL agent in the AP MLD create proper dynamic traffic steering for STA MLDs in an MLO-enabled Wi-Fi network.

V. Simulation

In this section, we evaluate the performance of the proposed model-free dynamic traffic steering methods on reward convergence, average network throughput, and load balancing ability in an IEEE 802.11be Wi-Fi network boasting MLO with the STR functionality under an ns-3 based simulation environment. For the proposed DRL approach, we apply the standard SAC implementation of the Stable Baselines3 library [13] to the DRL agent in the AP MLD, and employ the ns3-ai software framework [14] for the communication between the DRL agent and the ns-3 based simulation environment.

For average network throughput performance, we compare the proposed methods with the following baseline methods:

- **Round robin (RR) traffic steering**: The unallocated UL packets in an STA MLD are steered to available links in an RR fashion over each UL transmission window.
- **Min-queue (MQ) traffic steering**: The unallocated UL packets in an STA MLD are steered to the available link with minimum queue over each UL transmission window.

A. Parameter Settings

The IEEE 802.11be Wi-Fi network is configured in ns-3 as follows. We consider a Wi-Fi network which consists of a single AP MLD and a varying number M of STA MLDs with UL traffic. Besides, we consider $L = 3$ available links at 2.4, 5, and 6 GHz frequency bands with respective channel bandwidths of 20, 40, and 80 MHz. In each available link, we consider the presence of non-MLD overlapping basic service set (OBSS) that may interfere on the link. The number of OBSS's on each available link varies across simulations. We denote by O_l the number of OBSS's on the lth available link, $l = 1, 2, ..., L$. The summary of these and additional Wi-Fi network parameters for simulations is collected in Table I.

TABLE I
Wi-Fi Network Parameter Settings

Parameter	Value
(# AP MLD, # STA MLD)	(1, M)
Frequency band	2.4, 5, 6 GHz
Channel bandwidth	20, 40, 80 MHz
# OBSS on the lth available link	O_l
Duration of a UL transmission window	$\tau_{\text{UL}} = 10$ ms
# UL transmission window	$T = 50$
MAC protocol data unit (MPDU) payload size	256 bytes
Channel model	IEEE 802.11be indoor
(AP Tx power, STA Tx power)	(20 dBm, 17 dBm)
AP/STA noise figure	7 dB
Multiple-input multiple-output (MIMO)	2×2

The minimum and maximum values of weight in the proposed adaptive scoring heuristic (AS) algorithm are set as $(w_{\text{min}}, w_{\text{max}}) = (0.3, 3)$. The proposed SAC-based DRL (SAC) approach is configured as follows. With a learning rate of 0.0005, the actor policy and critic Q-networks are parameterized with multi-layer perceptrons with 3 layers of respective sizes of 256, 128, and 64 neurons. The summary of these and additional SAC parameters for simulations is collected in Table II.

TABLE II
SAC Parameter Settings

Parameter	Value
Actor/critic learning rate	0.0005
Actor/critic output layer activation function	Softmax
# neuron	[256,128,64]
Optimizer	Adam with step size 0.002
(γ, ρ)	(0.99,0.005)

B. Simulation Results

First, we evaluate the reward convergence of the proposed SAC approach under the configuration $(M, O_1, O_2, O_3) = (5, 2, 1, 0)$ with simulation results shown in Fig. 2. It is observed that the reward converges rapidly after around eight episodes, which confirms the robust adaptability of the proposed SAC approach for MLO dynamic traffic steering.

Next, we compare the proposed methods against the baseline methods in terms of average network throughput under two different configurations $(M, O_1, O_2, O_3) = (3, 2, 1, 1)$ and $(M, O_1, O_2, O_3) = (5, 2, 1, 0)$. The simulation results are shown in Fig. 3, demonstrating that the proposed methods not only outperform the baseline methods but also work well under different configurations.

Finally, we inspect the load balancing ability of the proposed methods with the evolving traffic steering portion over the

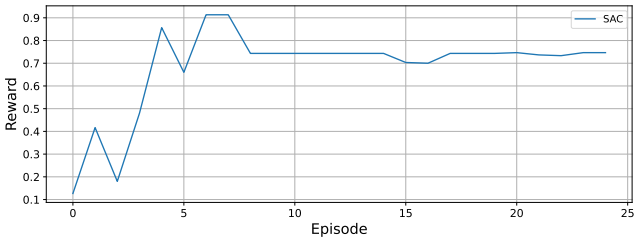


Fig. 2. Reward convergence of proposed SAC approach

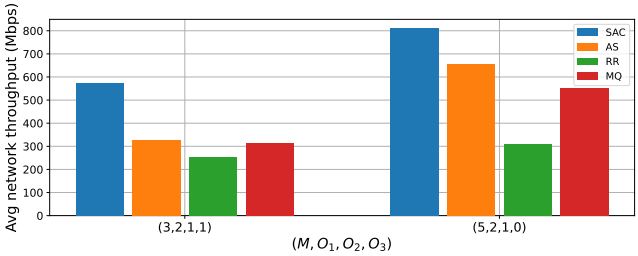
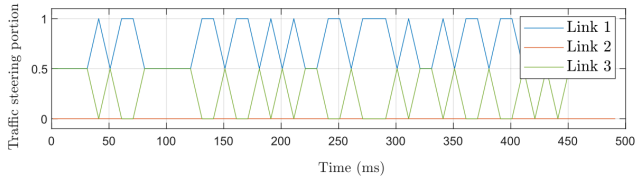
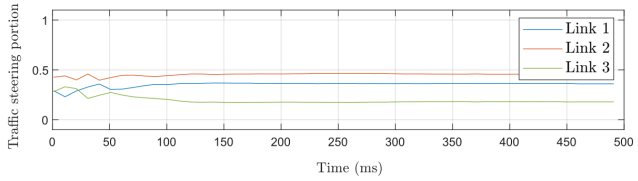


Fig. 3. Average network throughput under different configurations

time duration of 500 ms ($T \cdot \tau_{\text{UL}}$) under the configuration $(M, O_1, O_2, O_3) = (3, 2, 1, 1)$. Figs. 4(a) and 4(b) show the evolution of traffic steering portion in an STA MLD (randomly selected from three which demonstrate similar trends) over time with the proposed AS algorithm and SAC approach, respectively. From Figs. 4(a) and 4(b), it is inferred that the proposed SAC approach strikes a decent load balancing among all available links while the proposed AS algorithm tends to rely on some available links. Furthermore, the evolution of traffic steering portion with the proposed SAC approach is more smooth than that with the proposed AS algorithm.



(a) Proposed AS algorithm



(b) Proposed SAC approach

Fig. 4. Load balancing with evolving traffic steering portion over time

VI. Conclusion

In this paper, we consider STR-enabled MLO and propose two application-agnostic model-free dynamic traffic steering methods, where the AP MLD dynamically determines a flexible traffic steering portion for STA MLDs according to crucial factors such as SNR, number of allocated UL packets, and channel busy time. First, we develop an adaptive scoring heuristic algorithm based on a normalized weighted score and analyze its lightweight computational complexity. Second, we address the problem from the DRL perspective and develop a DRL approach based on standard SAC with robust adaptability. Simulation results exhibit the efficacy of the proposed model-free dynamic traffic steering methods in terms of reward convergence and average network throughput. Moreover, we show the different load balancing ability of the proposed model-free dynamic traffic steering methods.

Acknowledgment

This work was supported in part by the Wayne J. Holman Chair and the EVP for Research at Georgia Tech.

References

[1] A. Garcia-Rodriguez, D. López-Pérez, L. Galati-Giordano, and G. Geraci, "IEEE 802.11be: Wi-Fi 7 Strikes Back," *IEEE Communications Magazine*, vol. 59, no. 4, pp. 102–108, 2021.

[2] Á. López-Raventós and B. Bellalta, "Multi-link operation in IEEE 802.11 be WLANs," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 94–100, 2022.

[3] G. Lacalle, I. Val, O. Seijo, M. Mendicute, D. Cavalcanti, and J. Perez-Ramirez, "Analysis of Latency and Reliability Improvement with Multi-Link Operation over 802.11," in *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, 2021, pp. 1–7.

[4] M. Carrascosa, G. Geraci, E. Knightly, and B. Bellalta, "An Experimental Study of Latency for IEEE 802.11be Multi-link Operation," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 2507–2512.

[5] M. Carrascosa-Zamacois, G. Geraci, E. Knightly, and B. Bellalta, "Wi-Fi Multi-Link Operation: An Experimental Study of Latency and Throughput," *IEEE/ACM Transactions on Networking*, 2023.

[6] N. Korolev, I. Levitsky, and E. Khorov, "Analyses of NSTR Multi-Link Operation in the Presence of Legacy Devices in an IEEE 802.11 be Network," in *2021 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2021, pp. 94–98.

[7] W. Murti and J.-H. Yun, "Multi-Link Operation with Enhanced Synchronous Channel Access in IEEE 802.11be Wireless LANs: Coexistence Issue and Solutions," *Sensors*, vol. 21, no. 23, 2021.

[8] W. Zhan, B. Wu, X. Sun, K. Xie, and X. Chen, "Fairness-Constrained Rate Optimization of Multi-Link Slotted Aloha," *IEEE Networking Letters*, vol. 5, no. 1, pp. 31–35, 2023.

[9] Á. López-Raventós and B. Bellalta, "Dynamic traffic allocation in IEEE 802.11 be multi-link WLANs," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1404–1408, 2022.

[10] P. E. Iturria-Rivera, M. Chenier, B. Herscovici, B. Kantarci, and M. Erol-Kantarci, "RL meets Multi-Link Operation in IEEE 802.11be: Multi-Headed Recurrent Soft-Actor Critic-based Traffic Allocation," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 4001–4006.

[11] P. Christodoulou, "Soft actor-critic for discrete action settings," *arXiv preprint arXiv:1910.07207*, 2019.

[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 1861–1870.

[13] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 12 348–12 355, 2021.

[14] H. Yin, P. Liu, K. Liu, L. Cao, L. Zhang, Y. Gao, and X. Hei, "Ns3-Ai: Fostering Artificial Intelligence Algorithms for Networking Research," in *Proceedings of the 2020 Workshop on Ns-3*, ser. WNS3 '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 57–64.