THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Reinforcement Learning-Based Near-Optimal Load Balancing for Heterogeneous LiFi WiFi Network

Citation for published version:
Ahmad, R, Soltani, MD, Safari, M & Srivastava, A 2022, 'Reinforcement Learning-Based Near-Optimal Load Balancing for Heterogeneous LiFi WiFi Network', *IEEE Systems Journal*, vol. 16, no. 2, pp. 3084 - 3095. https://doi.org/10.1109/JSYST.2021.3088302

Digital Object Identifier (DOI):
10.1109/JSYST.2021.3088302

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Systems Journal

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 24. Apr. 2024

Reinforcement Learning-based Near-Optimal Load Balancing for Heterogeneous LiFi WiFi network

Rizwana Ahmad*, Mohammad Dehghani Soltani†, Majid Safari†, Anand Srivastava§
*§Department of Electronics and Communication, IIIT-D, New Delhi, India
† Institute for Digital Communications, School of Engineering, The University of Edinburgh, Edinburgh, UK
Email: *rizwanaa@iiitd.ac.in, †M.Dehghani@ed.ac.uk,

***Abstract*—Owing to the non-overlapping spectrum, Light fidelity (LiFi) and WiFi technologies can coexist and form a heterogeneous LiFi WiFi network (HLWN). The performance of HLWN significantly depends upon the load balancing strategies. Since load balancing of HLWN is a non-convex mixed-integer nonlinear programming (MINLP) optimization problem, it is mathematically intractable, and therefore, the conventional optimization methods fail to provide an optimal global solution. Although an optimal solution can be obtained using the exhaustive search method, it would be computationally complex. Therefore, in this paper, a reinforcement learning (RL) based algorithm is explored for solving the load balancing problem for the downlink HLWN at reasonably low complexity and near optimal performance. We have proposed three different reward functions for RL; the first and second reward functions work toward maximizing average network throughput and user satisfaction, respectively. The third reward function is designed to maximize the long-term system throughput and ensure at least 50% user's satisfaction for all users. In order to study the effects of link aggregation on system performance, this work considers two different types of receiver schemes, namely, single access point (SAP) and link aggregation (LA) scheme. While the SAP allows the user to receive data only from a single AP, the LA scheme allows the user to receive data simultaneously from both LiFi and WiFi AP. This paper also includes effect of random orientation of the receiver device and handover overhead. Further, concepts of domain knowledge have been included in this work to reduce the computational complexity of the algorithm. The proposed system performance is compared with the two benchmarks: received signal strength (RSS) and exhaustive search based on the computational complexity, average system throughput, and user satisfaction. It is shown that the proposed RL scheme outperforms the RSS scheme in average system throughput and user satisfaction. The RL scheme with an appropriate reward function provides a matching performance to the exhaustive search at reasonably low complexity.**

***Index Terms*—Heterogeneous LiFi WiFi network (HLWN), Light Fidelity (LiFi), Load balancing (LB), Link aggregation, Reinforcement Learning (RL), mixed-integer nonlinear programming (MINLP), Reward Shaping.**

I. Introduction

Due to Covid 19 pandemic, a 20% increase in the network traffic has been observed in the telecommunication industry as compared to pre-pandemic era [1]. With the on-going situation, this network traffic is further likely to increase and to support this growing future data requirement, many researchers are investigating the visible light communication (VLC) for the indoor environment. In VLC, information is transmitted using light emitting diodes (LEDs), by modulating the intensity of light. At the receiver, a photodetector (PD) converts the received optical signal into equivalent electrical signal. VLC offers inherent security, cheaper installation, immunity against electromagnetic interference, and huge spectrum in unlicensed band. Furthermore, unlike radio frequency (RF) counterparts, VLC does not cause health hazards and it provides high spatial reuse [2, 3]. Light fidelity (LiFi) is the wireless networking extension of the point-to-point VLC, which is capable of supporting a fully networked, bidirectional, and high-speed wireless communication. The universal availability of LEDs, license-free deployment and data rate of Gbps order, makes LiFi an attractive and inexpensive choice for indoor communications [4]. However, since light-wave cannot penetrate through opaque objects, LiFi suffers from a major drawback of blockage. Hence, the LiFi user's throughput fluctuates spatially, this results into various coverage holes in an indoor LiFi environment. LiFi can support high data rates when the receiver is in direct line-of-sight (LoS), but as soon as the LoS connection is lost, the data rate drops significantly; in comparison, WiFi can support moderate data rates with more ubiquitous coverage.

LiFi and WiFi technologies can coexist together and complement each other because of their non overlapping spectrum. In [5], it has been shown that a hybrid LiFi WiFi network (HLWN) provides higher system throughput as compared to standalone LiFi or WiFi networks. An appropriately designed HLWN can support higher data rate, better user satisfaction, outage performance, and lower handover rates [2].

For HLWN, load balancing (LB) is challenging as LiFi's and WiFi's coverage areas overlap with each other and WiFi covers larger area but has lower capacity; this increases the complexity of access point (AP) selection process. If the conventional received signal strategy (RSS) is applied for AP assignment in HLWN, WiFi AP will be susceptible to overload, and the system would not be able to ensure the required quality of service (QoS). A central control unit is required for efficient load balancing in HLWN [2]. The problem of load balancing in HLWN is a mixed-integer nonlinear programming (MINLP) problem. Furthermore, it must be noted that the problem of throughput maximization in HLWN is neither concave nor convex in binary connection variable [6, 7], therefore, the conventional optimization algorithms fail to find a global optimum for this problem. Hence, researchers have started exploring machine learning based solution for the aforementioned problem. The next subsection discusses the relevant literature related to LB in HLWNs.

A. Related Work

Various methods have been applied to solve the problem of LB in HLWNs, they can be broadly classified into following categories: optimization [6], machine learning [7–11] and fuzzy logic [12]. Limited studies have explored the application of machine learning for solving the LB problem in HLWN. In [7], the authors have proposed a deep Q-network learning based algorithm to solve the joint optimization problem of bandwidth, power and user association in HLWNs. They have also incorporated the concept of idle APs and transfer learning in their work. However, in their work they have focused on the data rate maximization and ignored user satisfaction, which is an equally important metric for a network performance. Additionally, they have not considered the effect of handover overheads and carrier-sense multiple access with collision avoidance (CSMA/CA) medium access of WiFi, in their work. In [8], authors utilized the context information of asymmetric uplink and downlink performance requirements of traffic for making the decision of network selection. This work considered instantaneous uplink and downlink throughput as their reward function. In [9], authors presented multi-armed bandit based AP selection strategies for HLWN, they modified decision probability distribution based upon two different algorithms: exponential weights for exploration and exploitation and exponentially weighted algorithm with linear programming. In our previous work [10], we have applied reinforcement learning (RL) for AP assignment in HLWN and reported promising results, but we assumed a simplistic system model. We assumed time-division multiple access, and assumed that a user can connect to single AP, i.e. it can either connect to WiFi AP or LiFi AP at a time. Furthermore, we did not considered the effect of receiver orientation or handover overhead in our previous system model. However, in current work the effect of link aggregation and different reward functions is studied. Further, concept of domain knowledge has been exploited to reduce the complexity of the proposed system. Additionally, in current paper, we have considered a more realistic framework for modeling the HLWN with CSMA/CA and handover overhead.

All of the above mentioned works are based on the assumption that a user is allowed to connected to a single AP at a time, limited literature covers the aggregation of LiFi WiFi links in a HLWN [11, 13–16]. In [11], authors proposed the concept of responsive and anticipatory association, the associations were established based on users geo-locations and queue backlog states. Their objective was to find the optimal trade-off between the average system queue backlog and the average per-user throughput. In [13, 14], authors have implemented channel aggregation for HLWN, and demonstrated proof-of-concept by using state-of-the-art LiFi and WiFi frontends. Both of these works focused on practical demonstration and employed AP assignment based on the received signal strength and ignored the effect of AP overloading. In [15], authors have proposed an online two-timescale power allocation algorithm for users with multi-homing capability that allows the users to aggregate the resources from both RF and LiFi APs. However, they have not considered the effect of interference between LiFi APs. Further, they implemented Q-learning based methods and compared their results only with stand-alone LiFi and WiFi AP, whereas in the current work policy-gradient based method has been implemented. In [16], the authors utilized Lyapunov optimization function for determining the optimal scheduling based on queue lengths for achieving the desired throughput. The authors have practically validated the performance of proposed protocol by implementing it on a real-life prototype. Their focus was on queue length-based scheduling algorithm to achieve optimal throughput. On the other hand, the current proposed work focuses on optimal AP assignment for a higher average network throughput while ensuring a particular user satisfaction. There are various research gaps in the above-mentioned works. Firstly, these works did not compare their results with two benchmarks RSS and exhaustive search method which has been done in the current work. Secondly, the current work also compares the performance of the system without and with link aggregation which was missing in most of the previous studies. Thirdly, the current work considers a more realistic scenario with effect of handover overhead, interference between neighbouring LiFi APs, user mobility and receiver device orientation. Additionally, in the current work, three different reward functions for different objectives have been explored, whereas most of the existing research works have explored only single reward function. Finally, in order to reduce the complexity of RL algorithm the concept of domain knowledge has been exploited in the current work.

B. Motivation and Main Contributions

Motivated by earlier works to study link aggregation effect on HLWN performance, two different types of receiver, namely, single AP (SAP) and link aggregation (LA) receiver, are considered in this work. The SAP receiver will allow users to receive data from either a LiFi or WiFi AP, whereas the LA receiver will allow users to receive data simultaneously from both LiFi and WiFi APs. For LA user, it is assumed that the physical and medium access layer of WiFi receiver and LiFi receiver will work independently [14]. Furthermore, in literature, it has been proven that load balancing of HLWN is non-convex MINLP problem [6, 7] which is mathematically intractable, thus, the conventional optimization algorithms can not find the global optimum solution for this kind of problem. Therefore, to overcome this limitation of the conventional optimization algorithm, we have proposed a centralized RL algorithm to perform LB in a HLWN. The RL algorithm determines its actions based on online-learning from the HLWN environment. Moreover, the effect of different reward functions on the proposed algorithm's performance is also studied in this work. The concepts of domain knowledge have been exploited to reduce the algorithm's computational complexity. It has been shown that RL based LB in a HLWN with an appropriate reward function provides significantly improved performance.

The main contributions of this paper are summarized as follows:

- Two different types of receivers, namely SAP and LA, have been considered to study the effect of link aggregation on the performance of HLWNs. A user equipped with a SAP receiver device can receive data from a single AP at a time, whereas a LA receiver equipped user can receive data simultaneously from both LiFi and WiFi APs.

- We have proposed a centralized RL based algorithm for dynamic LB in HLWN, and to prove the generalization of the proposed RL algorithm: three different reward functions have been considered, as explained in III-3 and their effect have been investigated on the average sum throughput and user satisfaction.
- A more realistic framework with orientation-based random waypoint (ORWP) mobility model, CSMA/CA-based multi-user access, and handover overhead has been considered in this work.
- Furthermore, the concepts of domain knowledge have been utilized to reduce the observation and action space, which reduces the RL algorithm complexity.
- The proposed RL scheme is compared against the RSS and exhaustive search method, explained in section III-A. The results are presented in terms of computational complexity, average network throughput, and user satisfaction.

The rest of the paper is organized as follow: Section II describes the system model and Section III introduces the proposed RL based LB method. The performance evaluation and discussion are presented in Section IV, and the paper is concluded in Section V.

II. System Model

In this paper, a typical office room of $5 \times 5 \times 3$ m^3 with multi-user HLWN is considered, as shown in Fig. 1. The coverage area of WiFi AP is assumed to be around 10 m, whereas the coverage area of each LiFi AP is limited to few meters. A central controller (CC) unit is required for efficient utilization of the HLWN [2]. In this work, it is assumed that a CC is connected to both WiFi and LiFi APs through an error-free feedback link and it is responsible for making the load balancing decisions. Furthermore, in order to study the effect of link aggregation, two different types of receiver are considered in this work, (1) SAP scheme which allows each user to receive data from a single AP, and (2) LA scheme which allows user to receive data simultaneously from both LiFi and WiFi AP. It is important to note that, although LA receiver, allows the user to simultaneously receive data from both LiFi and WiFi AP, but the CC can still decide to connect a LA user to only a LiFi/WiFi AP in order to avoid unnecessary overloading on a particular AP. Furthermore, the users are assumed to be accessing the high definition (HD) videos from the internet, hence their require data rates are modeled as a Poisson process with the parameter value of 70 Mbps. Moreover, in order to have a more realistic framework, the ORWP mobility model and CSMA/CA [17] based multi-user access have been considered in this work.

In the HLWN, the set of users is denoted by $\mathbb{U} = \{\mu | \mu \in [1, N_{\text{u}}]\}$. The set of LiFi APs is denoted by $\mathbb{LAP} = \{\alpha | \alpha \in [1, N_{\text{AP-LiFi}}]\}$ and WiFi AP is denoted as $\mathbb{W}$. The complete AP set is given by $\mathbb{AP} = \{\mathbb{W}, \mathbb{LAP}\}$. The total number of APs and users present in the system are represented by $N_{\text{AP}} = N_{\text{AP-LiFi}} + N_{\text{AP-WiFi}}$ and N_{u}, respectively. The LA user is capable of receiving data from both LiFi and WiFi AP simultaneously, therefore, let $\alpha_1 \in \{0, \mathbb{W}\}$ and $\alpha_2 \in \{0, \mathbb{LAP}\}$ indicate the LA users connection with WiFi and LiFi AP respectively. The value $\alpha_1 = 0$ means the user is not connected to the WiFi AP, similarly, $\alpha_2 = 0$ denotes the user is not connected to any LiFi AP.

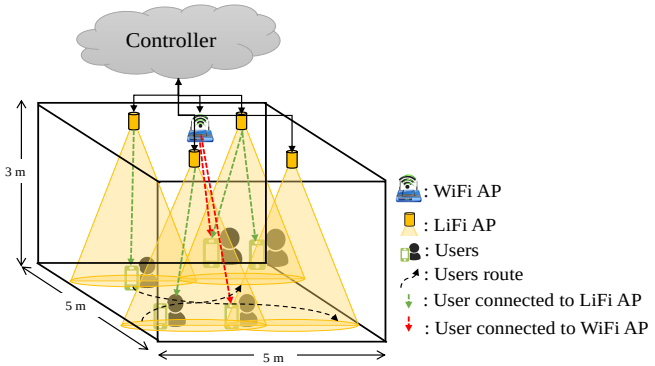


Fig. 1: Schematic diagram of a HLWN.

A. LiFi Channel Model

The signal-to-noise ratio (SNR) for the user μ connected to LiFi AP α is represented as $SNR_{\mu,\alpha}$, and can be expressed as:

$$SNR_{\mu,\alpha} = \frac{(H_{\text{LiFi}(\mu,\alpha)} P_{\text{opt}} \mathcal{R}_{\text{PD}})^2}{N_{\text{LiFi}} B_{\text{LiFi}}}, \tag{1}$$

where $H_{\text{LiFi}(\mu,\alpha)}$ is the channel gain between AP α and user μ, $\mathcal{R}_{\text{PD}}$ indicates photo receiver responsivity , P_{opt} represents transmitted optical power, N_{LiFi} is the LiFi noise power spectral density (PSD), B_{LiFi} indicates the LiFi AP bandwidth. The optical channel gain has two components: LoS and non LoS (NLoS) i.e., $H_{\text{LiFi}} = H_{\text{LoS}} + H_{\text{NLoS}}$. The LoS channel gain is defined as [10]:

$$H_{\text{LoS}} = \frac{(m+1) A_{\text{PD}}}{2\pi d^2} \cos(\phi) g_{\text{f}} g_{\text{c}}(\psi) \cos(\psi), \tag{2}$$

where m represents the Lambertian order of LED g_{f} and g_{c} defines gain of the optical and concentrator, A_{PD} indicates PD physical area, and d is the distance between LiFi AP and user. The ψ is the PD field of view (FOV). The NLoS channel gain is given by [10]:

$$H_{\text{NLoS}} = \frac{\rho A_{\text{PD}} e^{j2\pi f \Delta T}}{A_{\text{room}}(1-\rho)(1+j\frac{f}{f_c})}, \tag{3}$$

where ρ denote walls reflectivity, A_{room} indicates room area, ΔT is the delay between the LoS and diffused signals, and f_c represents cut-off frequency.

As all LiFi APs operate on the same frequency, they cause interference among each other. It is important to note that the interference term depends upon the actively transmitting AP that can be obtained by using the binary association variable $g_{\mu,\alpha}$, which is defined as:

$$g_{\mu,\alpha} = \begin{cases} 1, & \text{user } \mu \text{ is connected to AP } \alpha \\ 0, & \text{user } \mu \text{ is not connected to AP } \alpha \end{cases} \tag{4}$$

The AP β and AP α interfere with each other, if and only if both $g_{\mu,\alpha}$ and $g_{\mu',\beta}$ are set to 1, i.e. both the APs are transmitting to some users at the same time. In order to model this, we have defined a variable $I(g_{\mu,\beta}) = (1 - \prod_{\mu' \in \mathbb{U} \setminus \{\mu\}} (1 - g_{\mu',\beta}))$

TABLE I: LiFi channel parameters

Channel Parameter	Symbol	Value
Height difference between the user and AP	h	2 m
PD's Area	A_{PD}	1 cm^2
Optical filter's gain	g_{f}	1
Half intensity radiation angle	$\theta_{1/2}$	60°
PD's FOV	Ψ	60°
Responsivity	$\mathcal{R}_{\text{PD}}$	0.53 A/W
Reflection coefficient	ρ	0.8
LiFi AP's optical power	P_{opt}	3 Watt
LiFi AP's bandwidth	B_{LiFi}	40 MHz
LiFi noise PSD	N_{LiFi}	10^{-21} A^2/Hz

TABLE II: WiFi channel parameters

Channel Parameter	Symbol	Value
Breakpoint distance	d_{BP}	5 m
Shadowing loss	X_{SF}	3 dB
Central carrier frequency	f_{c}	2.4 GHz
WiFi AP's transmit Power	P_{WiFi}	20 dBm
WiFi AP's bandwidth	B_{WiFi}	20 MHz
WiFi noise PSD	N_{WiFi}	-174 dBm/Hz

and defined the signal-to-interference-noise ratio (SINR) between user μ and LiFi AP α as:

$$\text{SINR}_{\mu,\alpha} = \frac{(H_{\text{LiFi}(\mu,\alpha)} P_{\text{opt}} \mathcal{R}_{\text{PD}})^2}{N_{\text{LiFi}} B_{\text{LiFi}} + \sum_{\beta \in \mathbb{AP} \setminus \{\alpha\}} I(g_{\mu,\beta})(H_{\text{LiFi}(\mu,\beta)} P_{\text{opt}} \mathcal{R}_{\text{PD}})^2} \tag{5}$$

where $H_{\text{LiFi}(\mu,\beta)}$ is the channel gain between interfering LiFi APs β and the user μ and $\mathbb{AP}\backslash\{\alpha\}$ represents a set that includes all elements of set $\mathbb{AP}$ excluding element α. The lower bound on achievable data rate of the user μ connected to LiFi AP α can be calculated using [10]:

$$dr_{\mu,\alpha} = \frac{B_{\text{LiFi}}}{2} \log_2 \left(1 + \left(\frac{6}{\pi e}\right) SINR_{\mu,\alpha}\right). \tag{6}$$

The simulation parameters used for LiFi channel are summarized in Table I, which is same as [10, 12].

B. WiFi Channel Model

The SNR for user μ connected to WiFi AP α_1 is given by:

$$SNR_{\mu,\alpha_1}(f) = \frac{|G_{\mu,\alpha_1}|^2(f) P_{\text{T}}}{N_{\text{WiFi}} B_{\text{WiFi}}}, \tag{7}$$

where $G_(\mu, \alpha_1)(f)$ represents WiFi channel gain, P_{T} indicates transmitted power, N_{WiFi} denotes PSD of noise in WiFi, and B_{WiFi} is the bandwidth of WiFi AP. The WiFi channel gain, $G_(\mu, \alpha_1)(f)$ is given by [10]:

$$G_{\mu,\alpha_1}(f) = \sqrt{10^{\frac{-L(d)}{10}}} h_{\text{r}}, \tag{8}$$

where f indicates the carrier frequency, h_{r} represents the small-scale fading gain which follows independent identical Rayleigh distribution with 2.46 dB average power. The $L(d)$ denotes the large-scale fading loss and it is given as [10]:

$$L(d) = \begin{cases} L_{\text{FS}}(d) + X_{\text{SF}}, & d < d_{\text{BP}} \\ L_{\text{FS}}(d_{\text{BP}}) + 35 \log(\frac{d}{d_{\text{BP}}}) + X_{\text{SF}}, & d \geq d_{\text{BP}} \end{cases}, \tag{9}$$

where, d represents distance between user μ and WiFi AP α_1, L_{FS} denotes the free space loss, d_{BP} indicates breakpoint distance and X_{SF} refers to the shadowing loss. The free space loss can be calculated by $L_{\text{FS}}(d) = 20 \log_{10}(d) + 20 \log_{10}(f) - 147.5$ (dB). As the system model consist of single WiFi AP, there will be no interference for WiFi users The achievable data rate between WiFi AP α_1 and user μ , can be calculated using:

$$dr_{\mu,\alpha_1} = B_{\text{WiFi}} \log_2(1 + SNR_{\mu,\alpha_1}). \tag{10}$$

The WiFi channel parameters used in simulation are stated in Table II, which is same as [10, 12].

C. Orientation based Random Waypoint Mobility Model

Generally, most studies consider random way point (RWP) model for mobility. In RWP, a user pick a random destination and travels at a constant speed towards that destination. Once the user arrives at the destination, the user selects a new destination and starts moving in that direction at a constant speed and this process continues. However, in case of LiFi users, the receiver device orientation plays a crucial role. Therefore, in this work we have considered Orientation based RWP (ORWP) model which was initially proposed in [18] and developed in [19] and [20]. In fact, the ORWP considers the orientation of receiver device while the users move. A correlated Gaussian random process should be generated for the polar angle during the movement of users. The parameters of the ORWP model are chosen from [18]. The ORWP mobility model has been used for the first time in the hybrid LiFi and WiFi networks in [21] to assess the performance of the hybrid system more realisticly and support dynamic load balancing for mobile users. In this study, we considered pause time in the simulation of ORWP which is ignored in [21]. Users may tend to stop for a while at each destination and then continue their movement. We assume pause time at each destination follow an exponential distribution with a mean value of 10 seconds [20].

D. Handover

During a network handover, an overhead occurs that causes a drop in the average data rate of the user involved in the handover. In order to model this reduction in data rate, handover efficiency was introduced in [6] which is defined as the fraction of overhead time to actual transmission time. However, it is important to note that the exact handover efficiency can not be calculated [21]. Therefore, an average handover efficiency is used to estimate the negative effect of handover on users' data rate.

In a HLWN environment, two types of handover exists, namely, vertical and horizontal handover. When a SAP user moves from LiFi AP to WiFi AP or vice-versa, it is termed as vertical handover (VHO) and when a user moves from LiFi AP to another LiFi AP, it is termed as horizontal handover (HHO). Assuming that user μ was previously served by AP α^{t-1} and is now being served by AP α^t. For SAP user, the estimated handover efficiency can be modelled as:

$$\eta_{\text{SAP}}(t) = \begin{cases} 1, & \alpha^t = \alpha^{t-1} \ \ \forall \ \ \alpha^t, \alpha^{t-1} \in \mathbb{AP} \\ \eta_{0,\text{HHO}}, & \alpha^t \neq \alpha^{t-1} \ \ \forall \ \ \alpha^t, \alpha^{t-1} \in \mathbb{LAP} \\ \eta_{0,\text{VHO}}, & \text{otherwise.} \end{cases}.$$

where, $\eta_{0,\text{HHO}}$ and $\eta_{0,\text{VHO}}$ denote average handover efficiency for HHO and VHO, respectively. Typically, the HHO

require a lower amount of overhead and they occur faster in comparison to the VHO. The reason is that the HHO happen among the same domain using the same wireless technology whereas VHO fall among different technologies, which in this work is between LiFi and WiFi. Thus, a higher value of $\eta_{0,\text{HHO}}$ =0.9 is chosen in comparison to the $\eta_{0,\text{VHO}}$ = 0.6 [21]. However, it may please be noted that the choice of these values does not affect the generality of our proposed algorithm.

On the other hand, a LA user can either connect to both LiFi and WiFi AP, or connect to only one of them, therefore, the modelling of VHO will be different from SAP. In this paper, VHO for LA user is modelled as handover from both LiFi and WiFi connection to either LiFi or WiFi and vice-versa. Therefore, the estimated handover efficiency for LA user, can be modelled as:

$$\eta_{\text{LA}}(t) = \begin{cases} 1, & \alpha_1^t = \alpha_1^{t-1} \text{ and } \alpha_2^t = \alpha_2^{t-1}. \\ \eta_{0,\text{HHO}}, & \alpha_1^t = \alpha_1^{t-1} \text{ and } \alpha_2^t \neq \alpha_2^{t-1}. \\ \eta_{0,\text{VHO}}, & otherwise \end{cases}$$
$$\forall \quad \alpha_1 \in \mathbb{W}, \alpha_2 \in \mathbb{LAP}.$$

III. Proposed Reinforcement-learning Load Balancing Methods

RL is a promising machine learning approach, an RL agent is capable of providing an optimal solution (policy) without the exact knowledge of the underlying mathematical model. RL agents directly learns its policy based on the interaction with the environment, the communication happens between agent and environment in terms of *action* and *reward*. The ultimate goal of RL agent is to determine a stochastic policy, which maps states to a probability distribution over actions, in order to maximize the cumulative *reward* [10]. Fig. 2, shows the application of RL for LB in HLWN. RL works based on three vectors, namely, *state, action*, and *reward*. The *state* vector defines the present status of the hybrid LiFi WiFi environment. The *action* vector defines the action of AP assignment taken by the RL agent, after observing the present status of the environment. The *reward* vector defines the reward received by the system after an action is taken by the system. Let S and A represent the state and action space, respectively. At a given time step $t \geq 0$, an agent will be in state $s_t \in S$, it will take an action $a_t \in A$, and will receives a corresponding instant reward $r_t = r(s_t, a_t) \in R$ and transits to a next state s_{t+1}. The CC trains the learning algorithm in order to obtain its policy $\pi_\theta(a_t|s_t)$ for AP association. This process is repeated and with each iteration, the system keeps moving toward the actions that provides maximum cumulative reward. At the end of the training process the optimal policy $\pi^*(a|s;\theta)$ is obtained, this policy can be used in the real time in order to predict appropriate AP assignment, calculate reward and next state, as explained in algorithm 1.

In this work, the state, action and reward are formulated as follows:

1) State Space S: The *state* vector S defines the current status of the hybrid LiFi WiFi environment, it provides necessary information to the agent to make its decision. In this work, as we are training the agent to determine the optimal AP assignment strategy, therefore we need to take into account the SNR between the user and various APs, which is defined by (1) and (7). Furthermore, as we are dealing with HLWN which

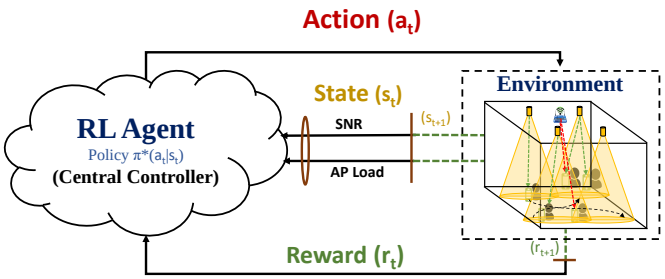


Fig. 2: RL for a HLWN.

is susceptible to AP overloading, we need to provide the information regarding load on a particular AP to the RL agent. Therefore, S is the set of continuous states, which includes:

- SNR between users and APs represented by a matrix $\mathbb{S}$
- Current load on each AP i.e., number of users connected to a particular AP represented by $\mathbb{L}_{\mathbb{AP}}$. The dimensions of $\mathbb{L}_{\mathbb{AP}}$ is $[N_{\text{AP}}]$

The dimensions of $\mathbb{S}$ depends upon the number of APs considered, conventionally the SNR matrix is considered between all the users and all the APs resulting into $\mathbb{S}$ dimension of $[N_{\text{u}} \times N_{\text{AP}}]$. However, based upon our domain knowledge, we have observed that in a standalone LiFi network, usually 2 APs on average provides a SNR difference of 10 dB. Therefore, instead of including the information about all the APs, it would be more efficient to transmit the SNR information between the user and two highest SNR providing APs to the controller. Therefore, by using this simple domain knowledge (DK) the dimension of the observation space can be reduced from $[N_{\text{AP}} + N_{\text{u}} \times N_{\text{AP}}]$ to $[N_{\text{AP}} + N_{\text{u}} \times (2+1)]$.

2) Action Space A: In this work, A is a finite set of multi-discrete actions. As we are considering the LA receivers, which allows the users to connect to both LiFi and WiFi AP at the same time. However, sometimes the simultaneous connection might not contribute towards a higher reward, this could happen due to AP overloading. In such cases, the controller can decide to transmit the information to LA user via single link only, therefore, converting LA into SAP receiver. Hence, the action space must have discrete values to indicate the standalone connection with LiFi or WiFi AP and simultaneous connection to both LiFi and WiFi AP as well. The action space for a particular user μ in our setup, can be defined as $A = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, where,

- $a_t = 0$ indicates the user is connected to WiFi AP,
- $a_t = 1, 2, 3, 4$ indicates that the user is connected to LiFi AP 1,2,3, or 4 respectively
- $a_t = 5, 6, 7, 8$ indicates that user is connected to both WiFi AP and LiFi AP 1, 2, 3 or 4 respectively.

Algorithm 1 RL based LB algorithm for HLWN

Input: Current state of HLWN, s_t and optimal policy $\pi^*_\theta(a|s)$
Output: Reward r_{t+1} and state s_{t+1}

1: Use $\pi^*_\theta(a_t|s_t)$ for given s_t, in order to predict a_t
2: Update the AP assignment in the environment based on a_t.
3: Based on new AP assignment calculate reward r_{t+1} and new state s_{t+1}.
4: **return** r_{t+1} and s_{t+1}

However, as we already discussed that two highest SNR APs are usually sufficient to serve the user demands. Therefore, by using this DK, we can reduce the action space. We can store the IDs of highest SNR providing APs at the user and can reduce the action space to $A = \{0, 1, 2, 3, 4\}$, where

- $a_t = 0$ indicates the user is connected to WiFi AP,
- $a_t = 1, 2$ indicates that the user is connected to highest or second highest SNR LiFi AP respectively
- $a_t = 3, 4$ indicates that user is connected to both WiFi AP and highest or second highest SNR LiFi AP respectively.

Inclusion of this simple DK, reduces both the action and observation space, which reduces the computational complexity and improve the convergence of the RL algorithm.

3) Reward: In this work, we have used three different reward functions (R_1, R_2, R_3) for three different objectives:

- R_1 : This reward is designed to maximize the long-term average network throughput. In case of SAP user, the immediate reward r_t, is given as:

$$r_t = \frac{\sum_{\mu \in \mathbb{U}} \sum_{\alpha \in \mathbb{AP}} (t_{\mu,\alpha})}{N_u}, \tag{11}$$

where, $t_{\mu,\alpha}$ is defined as:

$$t_{\mu,\alpha} = \begin{cases} \eta_{\text{SAP}} g_{\mu,\alpha} dr_{\mu,\alpha} k_{\mu,\alpha}, & \text{for SAP user.} \\ \eta_{\text{LA}} (g_{\mu,\alpha_1} dr_{\mu,\alpha_1} k_{\mu,\alpha_1} & \\ \quad + g_{\mu,\alpha_2} dr_{\mu,\alpha_2} k_{\mu,\alpha_2}), & \text{for LA user.} \end{cases} \tag{12}$$

where, $k_{\mu,\alpha}$ represents the time slot allocation between AP α and user μ, which is given as:

$$k_{\mu,\alpha} = \frac{1}{\sum_{\mu'} g_{\mu',\alpha}}, \quad s.t. \quad \mu' \in \mathbb{U}$$

It may be noted that the value of $k_{\mu,\alpha}$, depends only on the total number of users connected to AP α.

- R_2: This reward is designed to maximize the average long-term user satisfaction and the immediate reward r_t is defined as:

$$r_t = \frac{\sum_{\mu \in \mathbb{U}} \sum_{\alpha \in \mathbb{AP}} US_{\mu,\alpha} \times C_1}{N_u}, \tag{13}$$

where, C_1 scaling factor is included in order to avoid problem of local convergence and $US_{\mu,\alpha}$ is defined as:

$$US_{\mu,\alpha} = \begin{cases} \frac{\eta_{\text{SAP}} g_{\mu,\alpha} dr_{\mu,\alpha} k_{\mu,\alpha}}{R_\mu}, & \text{for SAP user} \\ \frac{1}{R_\mu} \eta_{\text{LA}} (g_{\mu,\alpha_1} dr_{\mu,\alpha_1} k_{\mu,\alpha_1} & \\ \quad + g_{\mu,\alpha_2} dr_{\mu,\alpha_2} k_{\mu,\alpha_2}), & \text{for LA user} \end{cases} \tag{14}$$

where, R_μ is the required data rate of user μ.

- R_3 : It is important to note that reward R_2 tries to maximize the average user satisfaction, which means even if a user is achieving very low user satisfaction and others are achieving high user satisfaction, the resultant average will be high. Therefore, the reward R_2 is incapable of ensuring the required QoS for every user. The reward R_3 is designed to maximize the long term average network throughput while ensuring 50% user satisfaction ($US_{\mu,\alpha} = 0.5$) for each user. A negative reward with appropriate scaling has been used to ensure 50% user satisfaction. The immediate reward r_t, is defined as:

$$r_t = \frac{\sum_{\mu \in \mathbb{U}} \sum_{\alpha \in \mathbb{AP}} Q_{\mu,\alpha}}{N_u}, \tag{15}$$

where, $Q_{\mu,\alpha}$ is defined as:

$$Q_{\mu,\alpha} = \begin{cases} -C_2 \times (1 - US_{\mu,\alpha}), & US_{\mu,\alpha} \leq 0.5. \\ C_1 \times US_{\mu,\alpha}, & otherwise. \end{cases} \tag{16}$$

where, $US_{\mu,\alpha}$ depends upon the receiver type and is defined by (14). Additionally, C_1 and C_2 scaling factors are included to avoid the problem of local convergence. Furthermore, value of $C_2 > C_1$ ensures that the condition of $US_{\mu,\alpha} \leq 0.5$ is highly discouraged.
The values C_1 and C_2 have been found intuitively by searching the space for values greater than 1. It was found that for $C_1 = 100$, the policy was able to converge to a global solution and beyond 100 there was no change in the system performance. Additionally, a higher value of C_2 means the agent will try more aggressively to avoid $US_{\mu,\alpha} \leq 0.5$ but this would penalise the average network throughput performance because there exists a trade-off between the average network throughput and user satisfaction. Therefore, $C_2 = 1000$ was set to provide a balanced performance in terms of the average network throughput and user satisfaction.

4) RL training Algorithm: The objective of the training process is to optimize the policy parameters θ in order to find the optimal policy, π^* which maximize the expected discounted return $\eta(\pi)$.

$$\pi^* = \underset{\pi}{\text{argmax}}(\eta(\pi)),$$

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right],$$

where $\gamma \in (0, 1)$ indicates the discount factor.

In this work, we have used a multi-layer perceptron (MLP) with parameters θ for the policy network and represented the policy as $\pi_\theta(a|s)$. For training of the policy network, we have used Trust Region Policy Optimization (TRPO) algorithm [22], which is a model-free policy gradient algorithm. TRPO supports good training stability [23] and guarantees monotonic improvement under certain assumptions [22]. TRPO enforces a Kullback Leibler (KL) divergence constraint on the size of update between the old and new policy at each iteration. The objective function for TRPO based policy optimization can be written as [22]:

$$\underset{\theta}{\text{maximize}} \, \mathbb{E}_{s \sim \rho_{\theta_{old}}, a \sim q} \left[\frac{\pi_\theta(a|s)}{q(a|s)} Q_{\theta_{old}}(s, a) \right], \tag{17}$$

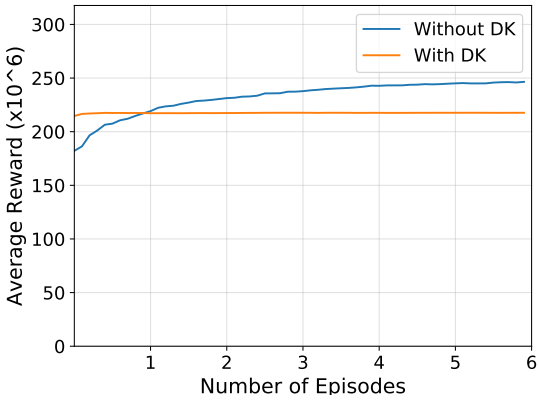
$$\text{subject to } \mathbb{E}_{s \sim \rho_{\theta_{old}}} \left[D_{\text{KL}}(\pi_{\theta_{old}}(.|s) || \pi_\theta(.|s)) \right] \leq \delta.$$

where, $Q_{\theta_{old}}$ is the state-action value function for policy $\pi_{\theta_{old}}$, $q(s|a)$ denotes the sampling distribution and δ is a tunable parameter. The state-action value function for a policy π, is represented by Q_π and is defined as:

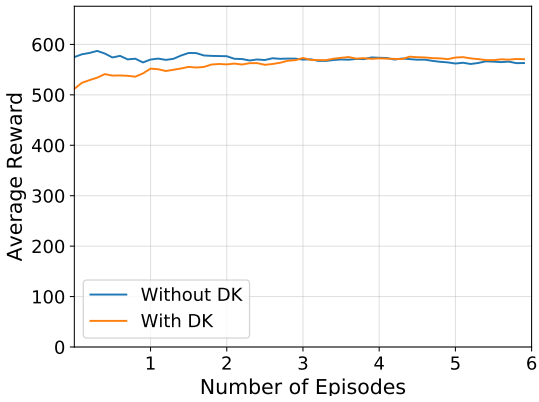
$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots} \left[\sum_{l=0}^{\infty} \gamma^t r(s_{t+l}, a_{t+l}) \right]. \tag{18}$$

The complete procedure of RL agent training based on TRPO is explained in [22]. In the next section, we will discuss about the training and convergence of the RL algorithm.

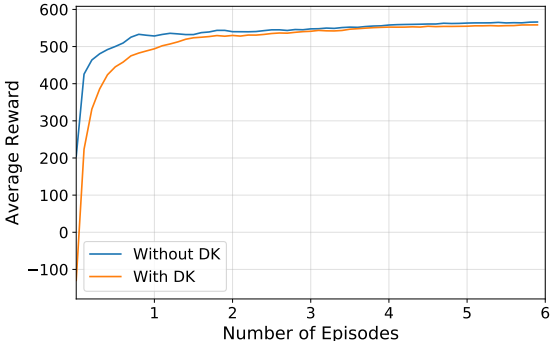
Training performance and convergence of RL: TRPO stabilizes the learning by imposing trust region constraints on the policy updation. TRPO being a model-free algorithm requires relatively lower hyper-parameter tuning, but its suffer with



(a) R_1 Reward



(b) R_2 Reward



(c) R_3 Reward

Fig. 3: Training performance and convergence of RL for different reward functions without (blue) and with (orange) DK.

high sample complexity [10]. In literature, various methods have been proposed to reduce the sample complexity of TRPO [24, 25]; in this work we have tried to reduce the complexity of TRPO by utilizing the DK which enables us to reduce the action and observation space. Therefore, reduces the overall complexity of the proposed system. The training performance of RL algorithm for three different reward functions with and without knowledge transfer is shown in Fig. 3. It can be observed that for all the reward functions the RL algorithm converges. From Fig. 3, it is clear that the RL agent converges to larger rewards when SNR from all the LiFi APs is considered, which is represented by blue colour curve. However, when only 2 best LiFi APs are considered the RL converges to a smaller average reward value indicated by orange colour curves. When R_1 is considered the difference in average reward with and without DK is significant, however, for R_2 reward both with and without DK converges to same values, as shown in Fig. 3 (b). Similarly, for R_3 reward both with and without DK converges to same value. It must be noted that only Fig. 3(c) has a negative value of average reward, this is due to the design of R_3 given in (15).

A. Other Load Balancing Methods

- **Received Signal Strength (RSS)** [12]: For a HLWN due to different physical receivers and bandwidth of LiFi and WiFi, the noise component observed at receiver is not uniform. Therefore, received signal strength does not fully represent the quality of channel. Hence, SNR must be used as the decision metric for RSS method in HLWN [10]. The objective function of the RSS method for a given user μ is defined as:

$$\max_{\alpha} \quad SNR_{\mu,\alpha} \qquad \text{s.t} \quad \alpha \in \mathbb{AP}. \tag{19}$$

where, $\mathbb{AP}$ is the set of APs including one WiFi and four LiFi APs and $SNR_{\mu,\alpha}$ represents the SNR values between μ user and α AP, which can be calculated by Eq. (7) and Eq. (1) for WiFi and LiFi APs, respectively. For SAP receiver, there will be single value of α given by (19). For LA receiver, the user will connect to two APs simultaneously, therefore, there will be α_1 and α_2, corresponding to highest SNR WiFi and LiFi AP, respectively. As there is only one WiFi AP present in the considered scenario, therefore α_1 = 1, indicating that user is always connected to WiFi AP. Another variable α_2, will give the value corresponding to highest SNR LiFi AP, which is defined as:

$$\alpha_2 = \max_{\alpha_2} \quad SNR_{\mu,\alpha_2} \qquad \text{s.t} \quad \alpha_2 \in \mathbb{LAP}. \tag{20}$$

- **Exhaustive search**: Exhaustive search also known as brute force search guarantees the best performance at the cost of high complexity. In line with the objective of the proposed RL scheme, the SAP users' objective function for exhaustive search with different rewards r_t is defined as:

$$\begin{aligned} &\max_{g_{\mu,\alpha} k_{\mu,\alpha}} \sum_{\mu\in\mathbb{U}} \sum_{\alpha\in\mathbb{AP}} (r_t), \\ &\text{s.t.} \sum_{\mu\in\mathbb{U}} (g_{\mu,\alpha} k_{\mu,\alpha}) = 1 \, \forall \alpha \in \mathbb{AP}, \\ &\sum_{\alpha\in AP} g_{\mu,\alpha} = 1 \, \forall \mu \in \mathbb{U}, \\ &g_{\mu,\alpha} \in \{0,1\}, k_{\mu,\alpha} \in [0,1], \forall \mu \in \mathbb{U}, \forall \alpha \in \mathbb{AP}. \end{aligned} \tag{21}$$

where, r_t is defined according to (11), (13), and (15) based on the corresponding rewards function. The first constraint ensures that the sum of time allocation of all users associated to one AP is 1 and the second constraint states that each user can get connected to only one AP at a time. Similarly, for LA receiver, user can connect to both AP's simultaneously, therefore, let $\alpha_1 \in \{0, \mathbb{W}\}$ and $\alpha_2 \in \{0, \mathbb{LAP}\}$ indicate the users connection with WiFi and LiFi APs respectively. For LA receiver, the objective

function for exhaustive search in line with proposed RL algorithm with different rewards r_t is defined as:

$$\begin{aligned} &\max_{g_{\mu,\alpha}k_{\mu,\alpha}} \sum_{\mu\in\mathbb{U}}\sum_{\alpha\in\mathbb{AP}}(r_t), \\ &\text{s.t.} \sum_{\mu\in\mathbb{U}}(g_{\mu,\alpha}k_{\mu,\alpha}) = 1\,\forall\alpha\in\mathbb{AP}, \\ &\sum_{\alpha_1\in\mathbb{W}} g_{\mu,\alpha_1} \le 1\,\forall\mu\in\mathbb{U}, \\ &\sum_{\alpha_2\in\mathbb{LAP}} g_{\mu,\alpha_2} \le 1\,\forall\mu\in\mathbb{U}, \\ &g_{\mu,\alpha}\in\{0,1\}, k_{\mu,\alpha}\in[0,1], \\ &\forall\mu\in\mathbb{U},\forall\alpha\in\mathbb{AP},\forall\alpha_1\in\mathbb{W},\forall\alpha_2\in\mathbb{LAP}. \end{aligned} \tag{22}$$

The first constraint is same as that of SAP user objective function. The second and third constraint states the condition that each user can get connected to maximum one WiFi and one LiFi AP at a time. The exhaustive search has been considered in order to provide the upper bound performance at the cost of higher complexity. The exhaustive search implementation has been made possible for this problem because of the room dimension, which restricts the number of users and APs, therefore limits the computational complexity to a reasonable value [10].

IV. Performance Evaluation and Discussion

We have considered a typical $5 \times 5 \times 3$ m^3 indoor space, with one WiFi and four LiFi APs, as shown in Fig. 1. It is assumed that the WiFi AP fully covers the room, whereas, the four LiFi APs partially covers the room. The focus of this work is to understand the effectiveness of RL based LB in HLWN, therefore, a simple scenario with four LiFi APs [10, 26] has been considered in this study. The proposed work is scalable to a larger room with more number of APs and users. Furthermore, two different types of receiver, i.e., SAP and LA schemes have been implemented in simulation. The effect of scheduling and reordering overhead is out of the scope of this paper, as that require the protocol design, which has been addressed in [16]. Moreover, in order to study the effect of different reward functions on system performance, three different rewards for optimising various system metrics, as explained in III-3 have been considered in this work. The simulation setup is coded in python 3.7 and MATLAB 2018. An Open AI Gym environment has been built from scratch for the HLWN. We have used stable-baseline GitHub repository [27] for RL algorithm (TRPO) and implemented ORWP [20]. The results reported are average over 200 episodes, and the values of system parameters are chosen in accordance with previously published works [10, 12, 21] and summarized in Table III.

The performance of the proposed RL with LA (RL-LA) method has been compared against exhaustive search with LA (Exh-LA), exhaustive search with SAP (Exh-SAP), RSS with LA (RSS-LA) and RSS with SAP (RSS-SAP), based on computational complexity, average network throughput and user satisfaction. Furthermore, this section also compares performance of RL-LA with DK (RL-LA-DK) against the Exh-LA with DK (EX-LA-DK), Exh-SAP with DK (Exh-SAP-DK), RSS-LA with DK (RSS-LA-DK) and RSS-SAP with DK (RSS-SAP-DK). The details regarding performance metrics are explained in next section.

TABLE III: System parameters

System Parameter	**Value**
Room dimension	$5 \times 5 \times 3$ m^3
Number of APs	4 LiFi + 1 WiFi
WiFi AP location	(2.5 m, 2.5 m)
LiFi AP locations	($\pm$1.25 m,$\pm$1.25 m)
User distribution	Uniform
User speed	1 m/s
User receiver	LA, SAP
Requested data rate, R_μ	Poisson with 70 Mbps
Gym environment	LiFi WiFi network
policy	MLP, 2 layers of 64
max KL divergence, δ	0.01
Discount factor, γ	0.9
Episode length, E	1000

A. Performance Metrics

In this work, the performance comparison is based on complexity, average network throughput and user satisfaction.

- The average network throughput (T) is calculated as:

$$T = \frac{\sum_{\mu\in\mathbb{U}}\sum_{\alpha\in\mathbb{AP}}(t_{\mu,\alpha})}{N_u}, \tag{23}$$

  where $t_{\mu,\alpha}$ represents the data rate of each user μ from the AP α, and can be calculated using (12)
- The users satisfaction $S_{\mu,\alpha}$ is defined as the ratio of data rate achieved by the user to the data rate required by that user, it can be expressed as:

$$S_{\mu,\alpha} = \min\{1, US_{\mu,\alpha}\}, \tag{24}$$

  where $US_{\mu,\alpha}$ is defined by (14). The user satisfaction ranges from 0 to 1, $S_{\mu,\alpha} = 1$ implies that the user has achieved the requested data rate.

B. Complexity Analysis

As RL-LA requires training and its convergence dependence on the state action space, and the RL algorithm. The training complexity of RL-LA cannot be directly compared with exhaustive search and RSS as these methods do not have a training phase. Therefore, in this paper, the training and convergence of RL is addressed separately in section III-4 and in this section, only run-time complexity of all methods is considered. In order to calculate complexity of RL-LA, it is important to note that the RLs' real time complexity in test phase is nothing but the complexity incurred in the forward pass of the trained policy network which in this work, is a MLP with 2 hidden layers. Let's assume, the number of neurons in each layer to be L_1 and L_2, respectively. The input and output layers will be defined based on the dimensions of observation and action space which are explained in section III. Therefore, the complexity of RL-LA is given by $\mathrm{O}((N_{\mathrm{AP}}N_{\mathrm{u}} + N_{\mathrm{AP}})L_1 + (L_1L_2) + (L_2N_{\mathrm{u}}))$.

The RSS-SAP method simply selects the AP with highest SNR value out of total APs (N_{AP}). Therefore, its complexity is $\mathrm{O}(N_{\mathrm{AP}}N_{\mathrm{u}})$ [12]. In RRS-LA method, user is always connected to WiFi AP and selects the highest SNR LiFi AP, therefore, its complexity is $\mathrm{O}(N_{\text{AP-LiFi}}N_{\mathrm{u}})$. The exhaustive search is computationally more expensive, since it looks for all possible

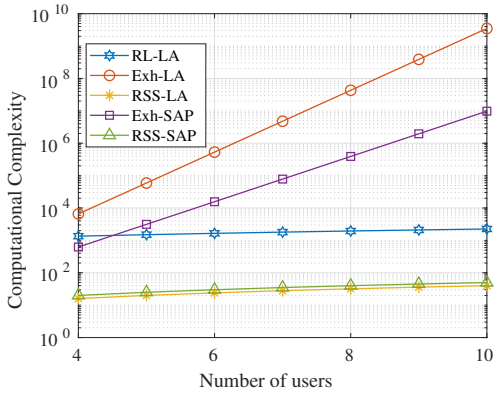


Fig. 4: Computational complexity of different schemes.

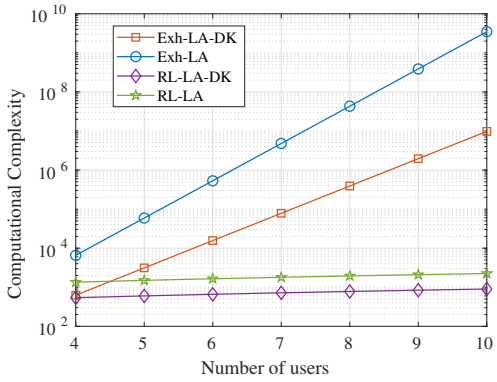


Fig. 5: Reduced computational complexity with DK.

connections between users and APs, therefore, the complexity of Exh-SAP is $\mathrm{O}((N_{\mathrm{AP}})^{N_{\mathrm{u}}})$. In case of Exh-LA, the complexity further increases to $\mathrm{O}((N_{\text{AP-WiFi}}+2N_{\text{AP-LiFi}})^{N_{\mathrm{u}}})$.

The complexity of various schemes with four LiFi and one WiFi AP, is illustrated in Fig. 4. It is clear that LA receiver has higher complexity as compared to SAP receiver. The application of DK can reduce the complexity of RL-LA and Exh-LA to $\mathrm{O}(((N_{\text{AP-WiFi}}+2)N_{\mathrm{u}}+(N_{\text{AP-WiFi}}+2))L_1+(L_1L_2)+(L_2N_{\mathrm{u}}))$ and $\mathrm{O}((N_{\text{AP-WiFi}}+4)^{N_{\mathrm{u}}})$, respectively. It is to be noted that RL run-time complexity can be reduced using neural network pruning [28] which is beyond the scope of this paper. The effect of DK on the computational complexity for Exh-LA-DK and RL-LA-DK is shown in Fig. 5.

C. Effect of Different Reward Functions

In this section the effect of different reward functions on average network throughput and user satisfaction has been presented.The average network throughput for different reward functions is summarized in Table IV. It can be observed that the performance of RSS-LA and RSS-SAP remains unchanged for different reward functions, this is due to the fact that for RSS-LA and RSS-SAP, the decision of AP assignment depends alone on the received signal strength and does not take into account the other factors. For all rewards, Exh-LA performs best followed by RL-LA in terms of average network throughput. The advantage of link aggregation can be clearly observed from Table IV. The RSS-LA provides an improvement of around 57 Mbps over RSS-SAP, similarly, Exh-LA provides an improvement of around 30 Mbps over Exh-SAP. The largest average network throughput of 235 Mbps

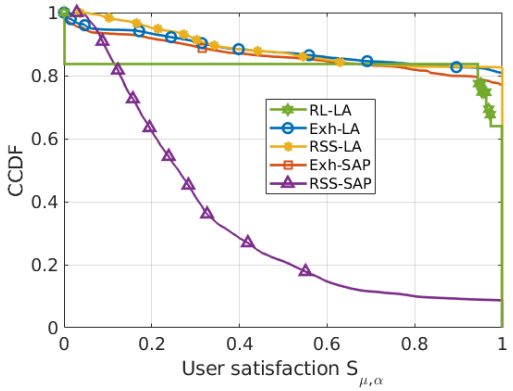
TABLE IV: Average network throughput (Mbps)

Reward	RSS-SAP	Exh-SAP	RSS-LA	Exh-LA	RL-LA
R1	40.66	190.83	97.60	235.81	220.90
R2	40.45	158.81	97.52	197.10	189.50
R3	40.86	174.81	97.80	218.90	215.90

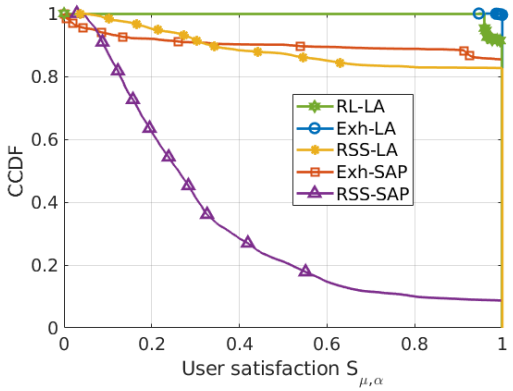
is achieved by Exh-LA for reward R_1, followed by RL-LA which is able to achieve 220 Mbps for reward R_1. The value of average network throughput in Exh-LA,RL-LA and Exh-SAP reduces for reward R_2, as R_2 focuses on maximization of average user satisfaction alone. The reward function R_3 provides a more balanced approach that ensures 50% user satisfaction and also tries to maximize the throughput. For reward R_3, the average network throughput improves over R_2, for Exh-LA, RL-LA and Exh-SAP. It is observed from Fig. 6 (a), that none of the schemes are able to ensure full user satisfaction for reward R_1, this is due to the fact that the reward R_1, is designed specifically to maximize the average network throughput. Therefore, we can see that a system with high average data rate does not guarantee a high QoS for users. The RSS-SAP receiver performs worst as it simply select one highest SNR AP for association. However, when a LA receiver is used, which allows the user to receive simultaneously from highest SNR LiFi and WiFi AP, the performance of RSS-LA improves significantly. As the reward function is focused on improvement of average network throughput, the performance of RL and exhaustive search with SAP and LA receiver suffers in-terms of user satisfaction. For the second reward function R_2, the results are shown in Fig. 6 (b). It can be seen that the performance of RSS-SAP and RSS-LA remains unchanged, as they are independent of the reward function. The reward function R_2 is specifically designed to maximize the average user satisfaction. For Exh-LA, there is significant improvement as it is able to provide full user satisfaction to all the users. Similar trend is observed for RL-LA, which is able to ensure full user satisfaction for 90% of the users and is able to ensure 96% user satisfaction for all users. There is also improvement in Exh-SAP, for reward function R_2, but this improvement is limited due to receiver restriction of single AP connection. The user satisfaction performance for R_3 reward function is shown in Fig. 6 (c). The results for RSS-SAP and RSS-LA remains unchanged. However, there is significant improvement in Exh-SAP. The reward R_3 ensures that all the users must achieve a user satisfaction of more than 50% and same can be observed from the Fig. 6 (c). For R_3 reward, RL-LA is able to provide full 98% user satisfaction to all the users. As Exh-LA, was already able to achieve full user satisfaction, therefore, no changes were observed in its performance.

D. Effect of Domain Knowledge

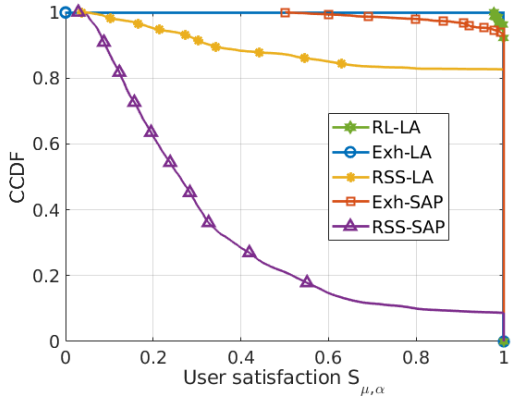
In this section, we present the results of domain knowledge transfer on various schemes with different reward functions. From the VLC domain knowledge, we understand that SNR information from two highest SNR LiFi APs is sufficient for making a decision of AP assignment, inclusion of this simple DK improves the convergence and reduces the computational complexity of proposed schemes. The Exh-LA-DK provides an improvement of around 40 Mbps over Exh-SAP-DK. The trend of average network throughput for various schemes with



(a) Reward R_1



(b) Reward R_2



(c) Reward R_3

Fig. 6: User satisfaction assessment for different reward functions.

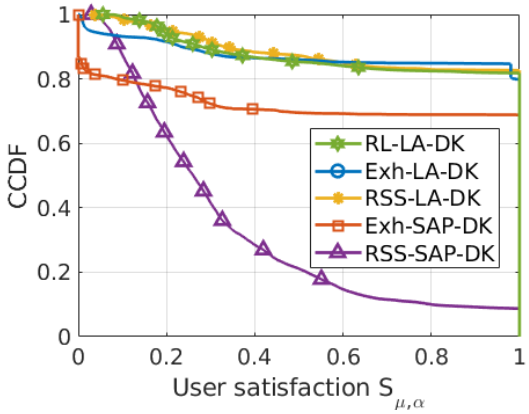
different rewards remains same as it is without the application of DK. However, in order to clearly understand the effect of DK, the values of Table V are compared with corresponding values from Table IV. It can be clearly observed that if only two highest SNR LiFi APs considered for decision making the average network throughput for Exh-LA-DK, RL-LA-DK and Exh-SAP-DK reduces as compared to when all the LiFi APs are considered. However, it is interesting to note that by the application of this simple DK reduces the gap between RL-LA-DK and Exh-LA-DK performance.

The user satisfaction for various schemes with DK, under different reward functions is illustrated in Fig. 7. As the reward R_1 focuses on maximization of average network throughput, the user satisfaction performance suffers, as shown in Fig. 7(a).

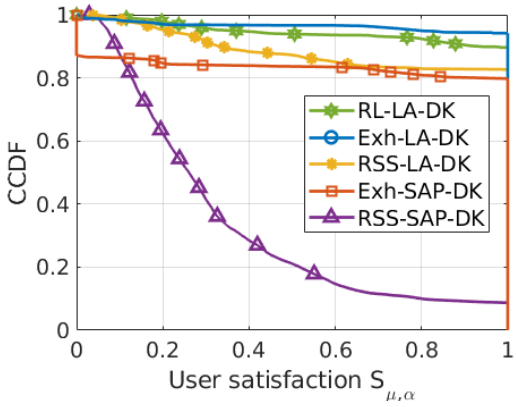
TABLE V: Average network throughput (Mbps) with DK.

Reward	RSS-SAP-DK	Exh-SAP-DK
R1	40.78	178.08
R2	40.15	143.76
R3	40.56	159.10

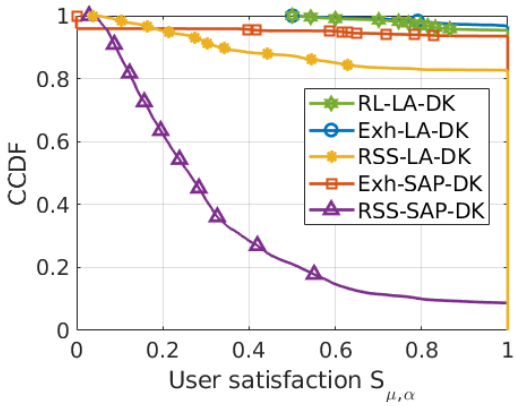
Reward	RSS-LA-DK	Exh-LA-DK	RL-LA-DK
R1	97.94	216.78	205.68
R2	97.17	180.97	173.20
R3	97.78	203.44	199.45



(a) Reward R_1



(b) Reward R_2



(c) Reward R_3

Fig. 7: User Satisfaction performance for different rewards with DK.

For reward R_2, the user satisfaction performance for Exh-LA-DK, RL-LA-DK and Exh-SAP-DK improves significantly.

From Fig. 7(b), it can be observed that Exh-LA-DK is able to provide full user satisfaction to around 95% of the users for reward R_2 as compared to 80% in case of reward R_1. Similarly, RL-LA-DK with R_2, is able to provide full user satisfaction to 90% of users as compared to 85% users when reward R_1 was used. The Exh-SAP-DK also observe 10% improvement in the number users achieving full user satisfaction for reward R_2 as compared to R_1. The user satisfaction performance of various schemes for reward R_3 with DK is shown in Fig. 7(c). It can be observed that reward R_3 provides best user satisfaction performance from individual users point of view. Even after application of DK and reduction of exploration space to only two highest SNR LiFi APs, Exh-LA-DK is able to provide full user satisfaction to 97% and RL-LA-DK is able to support around 96% of the users. The Exh-SAP-DK can ensure full user satisfaction to around 95% of the users. The application of reduced exploration space has a direct effect on the system performance which can be directly seen from Fig. 7. However, for reward R_3, a good user satisfaction and average network throughput can be achieved even while considering only two highest SNR LiFi APs. The application of DK reduces the system complexity significantly and its effect would be more prominent for a high density network deployment.

V. Conclusions

In this paper, RL based dynamic LB scheme for HLWNs is considered and three different rewards R_1,R_2, and R_3 have been investigated. For reward R_1, RL-LA is able to provide 106% improvement in average network throughput as compared to RSS-LA, but the user satisfaction was compromised. When reward R_2 is considered, RL-LA ensures full user satisfaction for 90% of the users and 96% user satisfaction for all users but the average network throughput was reduced. It is observed that RL-LA with reward R_3 provides a balanced system performance with high average network throughput (215.90 Mbps) and good user satisfaction (98%). Furthermore, we have also investigated the effects of link aggregation receivers on the system performance and it is observed that Exh-LA provides a minimum improvement of 23% over Exh-SAP in terms of average network throughput. Similarly, RSS-LA provides an improvement of around 57 Mbps over RSS-SAP. Therefore, we can conclude that LA significantly improves the system performance at the cost of increased complexity. The computational complexity for RL and exhaustive search increases quadratically and exponentially with the number of users. For the LA receiver scheme, the complexity further increases, which makes Exh-LA impractical for real-life scenarios. To reduce the computational complexity, this paper has introduced a concept from the domain knowledge transfer. It was observed that DK can significantly reduces the complexity at the cost of marginal performance degradation for Exh-LA-DK and RL-LA-DK. Overall, the RL-LA-DK with reward R_3 provides balanced average network throughput and user satisfaction performance which closely matches to the Exh-LA-DK and offers the advantage of low complexity.

Acknowledgment

M. D. Soltani and M. Safari acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) under program grant EP/S016570/1 'Terabit Bidirectional Multi-User Optical Wireless System (TOWS) for 6G LiFi'. Rizwana Ahmad would like to thank Intel India for PhD fellowship.

References

[1] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez *et al.*, "The lockdown effect: Implications of the covid-19 pandemic on internet traffic," in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 1–18.

[2] X. Wu, M. D. Soltani, L. Zhou, M. Safari, and H. Haas, "Hybrid LiFi and WiFi Networks: A Survey," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2021.

[3] D. N. Anwar and A. Srivastava, "Constellation design for single photodetector based CSK with probabilistic shaping and white color balance," *IEEE Access*, vol. 8, pp. 159 609–159 621, 2020.

[4] R. Ahmad, M. D. Soltani, M. Safari, and A. Srivastava, "Load Balancing of Hybrid LiFi WiFi Networks Using Reinforcement learning," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–6.

[5] H. Haas, L. Yin, C. Chen, S. Videv, D. Parol, E. Poves, H. Alshaer, and M. S. Islim, "Introduction to indoor networking concepts and challenges in LiFi," *Journal of Optical Communications and Networking*, vol. 12, no. 2, pp. A190–A203, 2020.

[6] Y. Wang, D. A. Basnayaka, X. Wu, and H. Haas, "Optimization of load balancing in hybrid LiFi/RF networks," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1708–1720, 2017.

[7] S. Shrivastava, B. Chen, C. Chen, H. Wang, and M. Dai, "Deep Q-network learning based downlink resource allocation for hybrid RF/VLC systems," *IEEE Access*, vol. 8, pp. 149 412–149 434, 2020.

[8] C. Wang, G. Wu, Z. Du *et al.*, "Reinforcement learning based network selection for hybrid VLC and RF systems," in *MATEC Web of Conferences*, vol. 173. EDP Sciences, 2018, p. 03014.

[9] J. Wang, C. Jiang, H. Zhang, X. Zhang, V. C. Leung, and L. Hanzo, "Learning-aided network association for hybrid indoor LiFi-WiFi systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3561–3574, 2017.

[10] R. Ahmad, M. D. Soltani, M. Safari, A. Srivastava, and A. Das, "Reinforcement learning based load balancing for hybrid LiFi WiFi networks," *IEEE Access*, vol. 8, pp. 132 273–132 284, 2020.

[11] R. Zhang, Y. Cui, H. Claussen, H. Haas, and L. Hanzo, "Anticipatory association for indoor visible light communications: Light, follow me!" *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2499–2510, 2018.

[12] X. Wu, M. Safari, and H. Haas, "Access point selection for hybrid Li-Fi and Wi-Fi networks," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5375–5385, Dec 2017.

[13] M. Ayyash, H. Elgala, A. Khreishah, V. Jungnickel, T. Little, S. Shao, M. Rahaim, D. Schulz, J. Hilt, and R. Freund, "Coexistence of WiFi and LiFi toward 5G: concepts, opportunities, and challenges," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 64–71, Feb. 2016.

[14] W. Zhang, L. Chen, X. Chen, Z. Yu, Z. Li, and W. Wang, "Design and realization of indoor VLC-Wi-Fi hybrid network," *Journal of Communications and Information Networks*, vol. 2, no. 4, pp. 75–87, Dec 2017.

[15] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid RF/VLC networks," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 443–447, 2019.

[16] Y. S. M. Pratama and K. W. Choi, "Bandwidth aggregation protocol and throughput-optimal scheduler for hybrid RF and visible light communication systems," *IEEE Access*, vol. 6, pp. 32 173–32 187, 2018.

[17] M. D. Soltani, X. Wu, M. Safari, and H. Haas, "Bidirectional User Throughput Maximization Based on Feedback Reduction in LiFi Networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3172–3186, Jul. 2018.

[18] M. D. Soltani, A. A. Purwita, Z. Zeng, H. Haas, and M. Safari, "Modeling the random orientation of mobile devices: Measurement, analysis and LiFi use case," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2157–2172, 2019.

[19] M. D. Soltani, M. A. Arfaoui, I. Tavakkolnia, A. Ghrayeb, M. Safari, C. M. Assi, M. O. Hasna, and H. Haas, "Bidirectional optical spatial modulation for mobile users: Toward a practical design for LiFi systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 9, pp. 2069–2086, 2019.

[20] M. D. Soltani, A. A. Purwita, Z. Zeng, C. Chen, H. Haas, and M. Safari, "An orientation-based random waypoint model for user mobility in wireless networks," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[21] Z. Zeng, M. D. Soltani, Y. Wang, X. Wu, and H. Haas, "Realistic indoor hybrid WiFi and OFDMA-based LiFi networks," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 2978–2991, 2020.

[22] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[23] T. Zhao, H. Hachiya, G. Niu, and M. Sugiyama, "Analysis and improvement of policy gradient estimation," in *Advances in Neural Information Processing Systems*, 2011, pp. 262–270.

[24] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel, "Model-ensemble trust-region policy optimization," in *International Conference on Learning Representations*, 2018.

[25] Y. Tang and S. Agrawal, "Boosting trust region policy optimization by normalizing flows policy," 2019.

[26] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 1, pp. 100–107, 2004.

[27] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Stable baselines," https://github.com/hill-a/stable-baselines, 2018.

[28] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *arXiv preprint arXiv:2003.03033*, 2020.

Rizwana Ahmad received her B.Tech. (2014) and M.Tech. (2016) degree from the Department of Electronics and Communication Engineering, Aligarh Muslim University, India. She is an Intel research fellow and working toward a Ph.D. degree in the Department of Electronics and Communication Engineering at IIIT-Delhi. Her research interests include visible light communication, light fidelity (LiFi) and coexistence of WiFi and LiFi networks.

Mohammad Dehghani Soltani received his Ph.D degree in electrical engineering from the University of Edinburgh, UK, in 2019. He is currently a Research Associate with the LiFi Research and Development Centre at the University of Edinburgh, funded by the British Engineering and Physical Sciences Research Council. His current research interests include mobility and handover management in wireless cellular networks, optical wireless communications, visible light communications and LiFi.

Majid Safari received Ph.D. in Electrical and Computer Engineering from the University of Waterloo, Canada in 2011. He is currently a Reader (Associate Professor) in the Institute for Digital Communications at the University of Edinburgh. Dr. Safari is currently an associate editor of IEEE Transactions on Communication and was the TPC co-chair of the 4th International Workshop on Optical Wireless Communication in 2015. His main research interest is the application of information theory and signal processing in optical communications including fiber-optic communication, free-space optical communication, visible light communication, and quantum communication.

Anand Srivastava did his Ph.D. (2002) from Indian Institute of Technology (IIT) Delhi, India. Before joining IIIT-Delhi in 2014, he was Dean and Professor in School of Computing and Electrical Engineering at IIT Mandi, HP, India (2012-2014), and also Adjunct Professor at IIT Delhi (2008-present). Prior to this, he worked with Alcatel-Lucent- Bell Labs (2009-2012), India, and Center for Development of Telematics (CDOT), India (1989-2008). His current research work is in the area of optical core and access networks, fiber-wireless (FiWi) architectures, visible light communications (VLC), optical signal processing, free space optical communications and energy aware optical networks.