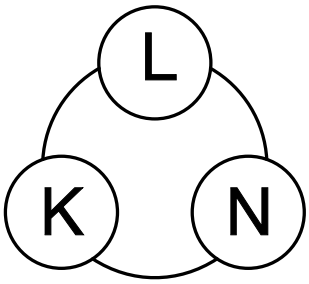

Technische Universität München
Lehrstuhl für Kommunikationsnetze
Prof. Dr.-Ing. Wolfgang Kellerer

Bachelor's thesis

Resource Allocation with Reinforcement Learning
to maximize rate in LiFi-WiFi Networks

Author:	Huber, Matias Robles
Address:	Kapuzinerstr. 12
	80337
	Munich, Germany
Matriculation Number:	03728435
Supervisor:	Vijayaraghavan, Hansini
Begin:	14. April 2023
End:	18. August 2023

With my signature below, I assert that the work in this thesis has been composed by myself independently and no source materials or aids other than those mentioned in the thesis have been used.

München, 18.08.2023	
Place, Date	Signature

München, 18.08.2023	
Place, Date	Signature

Abstract

This paper delves into the innovative concept of a hybrid LiFi/WiFi network (HLWN) that capitalizes on Light Fidelity (LiFi) and Wireless Fidelity (WiFi) technologies to augment indoor data rates and coverage. HLWN seamlessly integrates both technologies, leveraging LiFi's potential for higher data rates and its non-overlapping frequency spectrum, thus addressing the escalating bandwidth demands of future wireless systems. The cost-effective integration of LiFi into existing WiFi networks is a distinct advantage, facilitated by its compatibility with conventional light sources.

The core contribution of this study lies in the application of Reinforcement Learning (RL) models for optimizing resource allocation within the HLWN, with particular emphasis on managing horizontal handovers. The primary objective is to enhance user experience by minimizing unnecessary handovers and alleviating congestion issues. Through adept resource allocation decisions, the study effectively reduces excessive handovers, ensuring more seamless data transmission.

Additionally, the paper investigates the potential of incorporating mobility predictions into the RL framework, enhancing resource allocation foresight. Comprehensive examination of various RL models, including those with intricate action spaces, is an example the study's innovative research. Among tested RL algorithms, Advantage Actor-Critic (A2C) emerges as the optimal choice due to its stability and suitability for HLWN resource allocation, while Proximal Policy Optimization (PPO) demonstrates limitations.

Overall, this research unveils the potential of HLWNs and provides novel insights into the convergence of LiFi and WiFi technologies for resource-efficient communication systems.

Contents

Chapter 1

Introduction

1.1 Motivation

This paper discusses the concept of an HLWN, which combines LiFi and WiFi technologies to enhance data rates and coverage in indoor scenarios. In HLWN, LiFi and WiFi networks coexist and complement each other, as their frequency spectra do not overlap, and hence won't cause interference between different packages.

A heterogeneous LiFi-WiFi Network can provide various advantages compared to commonly used pure Radio Frequency (RF) Networks. LiFi offers the potential for significantly higher data rates, as visible light has a frequency range wider than RF, allowing greater bandwidths and a larger volume of data transmission. Additionally, the visible light frequency spectrum does not overlap with RF, making it viable to use both networks simultaneously, as well as giving access to relatively unused and uncongested frequencies; this is particularly important, as in 20 years the bandwidth demand of future wireless systems will exceed the RF capabilities 667 times[1].

Furthermore, implementing LiFi channels within already existing WiFi networks is a relatively cheap endeavor considering the already established system that can be exploited. LiFi technology can be easily integrated into existing light sources, with unnoticeable alteration to the lighting during data transmission.

A notable obstacle in applying a LiFi Network is the limited range of each LiFi Access Point (AP). Implementation of a *small cell concept* network, where WiFi APs prevent drastic drops in user data rates during LiFi handovers, is the solution investigated in this paper.

1.2 Contribution

This study presents a comprehensive exploration into the utilization of RL models for the optimal allocation of resources within an HLWN, specifically focusing on addressing horizontal handovers. The primary objective is to enhance the overall user experience by leveraging RL techniques to minimize unnecessary handovers and mitigate potential congestion issues arising from the overwhelming utilization of individual APs.

The RL model plays a pivotal role in achieving these goals. By making resource allocation decisions based on learned patterns, the model aids in reducing superfluous handovers, ensuring high user satisfaction and minimal interruption of data transmission. Furthermore, it acts as a safeguard against the congestion of APs and promotes a balanced network load distribution.

A proactive approach was undertaken to assess the potential incorporation of mobility predictions within the context of an HLWN. However, it is crucial to clarify that the proactive study does not aim to determine the viability of mobility predictions, instead, its main focus is to ascertain whether mobility predictions can introduce significant benefits to the resource allocation process. This entails investigating whether more intelligent and informed resource allocation decisions can be made with knowledge of future user positions.

Importantly, this paper delves into an in-depth exploration of various RL models, each possessing distinct complexities in their respective action spaces and algorithms. Notably, our study goes beyond the state-of-the-art by incorporating interference within the LiFi channels, which provided more reliable estimations of data rates within the simulations. The proactive approach referred to earlier involves integrating mobility prediction data into the RL framework, allowing for more foresighted decisions in resource allocation.

1.3 Organization

The remainder of this paper is organized as follows. Chapter 2 discusses the background of HLWN. This includes previous applications of RL and research in proactive resource allocation. In Chapter 3, the implementation of the RL in an HLWN is explained. The environment in which the experiments were applied is illustrated, as well as an in-depth description of the algorithms and formulas used, with a prominent explication of the *action* and *observation spaces*. The different models created, and their respective results, are analyzed in Chapter 4. And, lastly, in Chapter 5, a conclusion of the results, and an outlook on the overall research are provided.

Chapter 2

Background

A mobility-aware model would allow the control unit (CU) to preemptively avoid light-path blockages and bandwidth shortages. User mobility induces handovers in a system with stationary wireless channels, which causes degradation in user data rates, and may cause user connections to be blocked. A proactive model is considered a potential solution to this issue, as accurate predictions of future user positions can help APs reduce the probability of overloading by allocating resources in an anticipatory manner. This is particularly important in a network with LiFi APs, where handovers occur at a higher frequency.

LiFi networks can provide very high data transmission rates, however, are limited by the high frequency of handovers required to keep a user connected. A heterogeneous network is considered an ideal solution, as it uses already-established WiFi equipment, and provides high data rates with minimal interruptions. A LiFi system causes a larger frequency of user handovers due to the smaller coverage radii of 2m-3m [1] [2], and the handover efficiency is extremely influential to the user experience. Therefore even with minimal user movement, handovers may be required due to the intermittent light-path blockages. The impact of user mobility on the handover cost was extensively analyzed in [2], where horizontal handover (HHO) and vertical handover (VHO) were discussed. HHO occurs within a channel of a single wireless technology, whereas VHO occurs only in a hybrid channel. VHO however requires more complex computation. In this paper, an HHO network was analyzed, as the CU can connect users to a WiFi and LiFi AP simultaneously, and handovers would only occur between APs of the same technology.

A reinforcement learning (RL) algorithm is used in this paper, to find an optimal solution to maximize user satisfaction. An RL model is proposed due to its adaptability, and dynamic resource allocation. This paper applied two sets of RL

algorithms in the training process of the model:

1. Proximal Policy Optimization (PPO): PPO is a policy-based reinforcement learning algorithm. It belongs to the family of policy gradient methods, which focuses explicitly on exploration and is better suited for a continuous action space.

2. Advantage Actor-Critic (A2C): A2C combines the benefits of both policy-based and value-based methods. Policy gradient methods can handle continuous action spaces and encourage exploration, while value-based methods help stabilize learning and provide more accurate value estimates.

A policy-based algorithm focuses on discovering the relationship of the states and actions of the model, this promotes a more exploratory method of training. Whereas a value-based algorithm focuses on discovering the actions that output the best results, this reduces exploration but also promotes more stable training and lower computational costs due to its discrete nature; *Q-learning* is an example of a value-based algorithm.

2.1 LiFi Basics

A LiFi network is considered to be an ideal solution to the growing demand for frequency variation within current global wireless telecommunication systems. A major problem presented by such networks is dynamic handover, which causes interruptions in data transmission with users when alternating between APs. In [3], a soft handover solution was proposed, where a user remains connected to its previous AP before connecting to the next one; this demands more wireless transmission resources whilst providing a better user experience. Whilst connected to the LiFi or WiFi network, users would monitor both connections. In case of degradation of the signal strength of the current network, the user would send a signal to perform a VHO to the other network if it is optimal. In [3] HHOs were not implemented due to the homogeneous nature of the network; however, there was an extensive analysis of the possible decision schemes that may be used with VHO, which included mobility prediction and deep learning approaches.

Within an HLWN, it is important to consider signal interruptions caused by blockages between the line of sight (LOS) of the AP and the user. In [2], a fuzzy logic (FL) algorithm is used to optimize handovers, by considering user mobility and the impact of light-path blockages. Users were categorized into three groups (*LiFi Only*, *WiFi Only*, *LiFi/WiFi*), which determined which APs the user was allowed to connect to. Users within the heterogeneous network would connect to a WiFi AP in case of a light-path blockage and would return to a LiFi connection once

accessible. This logic was adapted and implemented within this paper. The control model used to compare with the RL models created is instructed to connect to a LiFi AP and WiFi AP if accessible, and the bandwidth is always equally distributed. Dissimilarly, in [2], there was no proactive approach, as the CU was allocating resources reactively, nor was RL implemented.

2.2 Proactivity

A network can improve the accumulated user experience over a specified period with a mobility-aware scheme. In [4], an average handover efficiency η_0 of 0.5 was used to induce data rate reduction for a single iteration when a handover occurred. The optimization function used was the aggregated data rate of all users within that time period, and through optimal resource allocation, overloading of APs and future deterioration to the user experience were sometimes avoidable. The proactive approach prioritized the accumulated data rates over the instantaneous user data rates. As an alternative to an exhaustive search algorithm, [4] used Lagrangian Multipliers to create a relaxation of the β-proportional fairness function (2.1), and no deep learning was involved. Furthermore, this model only considered a homogeneous network, so user connections to multiple APs were not an option.

$$\text{PF(user)} = \frac{\text{Instantaneous Rate(user)}^{1-\beta}}{1-\beta} \tag{2.1}$$

PF(user) refers to the fairness value for a specific user, Instantaneous Rate (user) is the immediate data rate of that user, and β is an adjustable value. When $\beta > 1$, the formula gives more weight to users with lower data rates, making the allocation fairer by prioritizing users with worse connections. When $0 < \beta < 1$, the formula gives more weight to users with higher data rates, focusing on performance rather than fairness.

In [5], various routing protocols were evaluated in determining the constantly changing topology in a mobile ad hoc network (MANET). These routing protocols were *reactive*, *proactive*, and *hybrid*:

1. Temporally-Ordered Routing Algorithm (TORA): TORA is a reactive routing protocol. The nodes within the system would initiate route discovery when trying to communicate with a destination, for which no previous route had been established. The route discovery helps find and integrate new connections.

2. Optimized-Link State Routing (OLSR): OLSR was the proactive example in the paper. It maintains routing for all nodes in the network updated for every

iteration. Each node sends topology control (TC) signals, with reachability and mobility information.

3. Zone Routing Protocol (ZRP): this hybrid protocol includes features from both other protocols in an attempt to leverage their advantages. This protocol would maintain routing information for a subset of the network, rather than maintaining the routing information throughout all iterations similar to a proactive protocol. The protocol would maintain proactive routing within specific zones for efficient localized communication.

A hybrid scheme is a very interesting approach, that can also be a potential solution to reduce the computational cost of a fully proactive approach. In this paper, only completely proactive and reactive models were analyzed, but when considering the potential computational cost of mobility prediction, a hybrid model may be more efficient and still provide exceptional user satisfaction.

In [6], a predictive scheduling scheme within diverse network environments is shown to optimize data transmissions. A *T-slot lookahead* scheduler algorithm is proposed, that is versatile and adaptable, making it suitable for various different networks; it aims to maximize resource usage while reducing collision and interference within the network. The *T-slot lookahead* is also used in this paper's reinforcement learning (RL) model, where a *window size* of 5 iterations was applied. The scheduler uses historic traffic data to predict user positions and proactively allocates resources to users dynamically, depending on the topology of the current environment. The scheduler also classifies the demand based on their requirements, as some traffic may demand higher bandwidths; in contrast, the model created in this paper considered all demands equal and only applied a penalty at significantly low data rates. This paper demonstrated the potential improvement of a proactive approach within a network and described the integral logic of such a scheme. However, these ideas were not implemented on a LiFi or WiFi network, nor was RL applied.

The predictive scheme used to determine user movement within an HLWN is a key component in a proactive approach. Various enhancements of a basic predictive channel reservation (PCR) scheme were discussed in [7], such as reservation pooling, integrating guard channels (GCs), applying a threshold distance (TD), and reservation queuing. These enhancements weren't directly applied to the RL models created in this paper, but, through correct training, these enhancements should be learned and applied optimally by the model. For instance, the TD can be learned by the model, allowing it to prioritize users with larger achievable SINR due to smaller distances, thereby making appropriate and more dynamic decisions regarding connecting to or receiving bandwidth from the AP.

2.3 Mobility Models

Various mobility models were also examined in [5], including Random Waypoint, Manhattan, and Group Mobility models. According to the results, the Random Waypoint model, which was the only mobility model used in this paper, was not an ideal candidate for all routing protocols. However, it is the most commonly used mobility model among MANET and HLWN simulations[4][3][2].

In [2], the heterogeneous environment was altered and tested with different mobility models. The different mobility models that were examined include:

1. *constant speed*: where a user's speed remains uniform between 0 and a max velocity
2. *varying speed*: based on the original random waypoint model (RWP), the user gradually changes speed when arriving at each waypoint
3. *varying speed with pausing*: a *varying speed* model with possible pauses between excursions

2.4 Model/Algorithm Logic

Deep reinforcement learning (DRL) in a heterogeneous network was explored in [8]. Where an adaptive handover mechanism was applied to the environment, allowing the model to dynamically allocate resources, in accordance with data rates and the number of users connected to an AP. In [8], the model was not directly responsible for the bandwidth proportions of each user, instead, the data rate was calculated by equally dividing the bandwidth among user devices (UDs).

$$R = \frac{B}{u_{i,t}} log_2(1 + \Gamma) \tag{2.2}$$

where $u_{i,t}$ represents the number of users connected to AP_i at time iteration t. In contrast, the RL model proposed by this paper defines the proportion of the bandwidth, as well as the UD connections to the APs. Another dissimilarity is the use of a proactive model, in order to anticipate potential excess demand on an AP. The DRL model in [8] distributed resources in a large-scale HLWNet, and, as compared in the paper, the traditional RL model, with a standard *Q-table* algorithm, would have a very low search efficiency due to the size of the observation space. The DQN algorithm used, had a deep convolutional neural network in place of a *Q-table*, provided an increase in the average downlink data rate by 13%.

An FL system comprises fuzzification, rule evaluation, de-fuzzification, and resource allocation; it was used in [9]. In this example however, a 3-state criterion

was applied, where the FL has three possible states of observation, that being *high quality*, *medium quality*, and *low quality* resource blocks. This defines the viability of allocating resources from a particular AP to a user. It elaborates on the differences between a traditional FL system, with binary decision-making, and the one used; the paper concluded there was a significant improvement in user satisfaction with the 3-state logic. In this paper, the precision of the *observation space* was not examined, but rather a continuous observation space is used in all models.

2.5 Summary

In summary, this paper introduces a novel proactive RL approach tailored for resource allocation within a heterogeneous LiFi-WiFi network impacted by user mobility. The proactive approach builds on traditional reactive strategies with a preemptive mindset, tackling challenges like light-path blockages and bandwidth shortages. It presents a departure from existing studies in several significant ways:

1. Proactive Resource Allocation: Unlike previous research that often employs reactive resource allocation strategies, this paper's approach proactively anticipates potential network issues by leveraging accurate predictions of user positions within the environment. This enables the CU to allocate resources preemptively, mitigating the negative effects of handovers and enhancing user experiences.

2. Logic of Resource Allocation: While previous studies have explored various optimization techniques, this paper harnesses RL algorithms, specifically PPO and A2C, to dynamically adapt resource allocation strategies based on real-time observations. Additionally, this paper also includes a comparative control model, with a simple logic base without an RL background, in order to provide a frame of reference in model evaluations.

3. LiFi-Specific Challenges: Traditional studies often overlook the potential implementation of LiFi within a network of data communication. In contrast, this paper's model is tailored to HLWNs, addressing the need for rapid handover management and optimized resource allocation within the LiFi channels.

In conclusion, this paper contributes a unique proactive RL solution to the resource allocation challenges in a LiFi-WiFi network with user mobility. By addressing LiFi-specific hurdles, and harnessing RL algorithms, this research builds on previous work by presenting an approach that potentially enhances network efficiency and user satisfaction.

Chapter 3

Implementation

3.1 Environment

Due to the inability of light waves to go through opaque solid walls and LiFi AP's limited radii, a simple LiFi network is mostly only suitable for smaller environments. The network examined was a confined indoor low-scale environment: a 5mx5mx3m room with 4 LiFi APs and a centralized WiFi AP. The number of users μ_{max} was 12 or 20, in order to analyze the results with more strenuous demands on the network. These users moved in accordance with an RWP, and this positioning was created by a separate script. The positioning tracked was that of the photodetector (PD) on each user UD, which was distributed around a height of $1.4m$. The sample time of each t time iteration was $0.01s$; with a user velocity of $1m/s$, this allowed a maximum $1cm$ of absolute distance from the previous position. At each service time, the control unit (CU) represented by the RL model, would be able to alter $\mu_{\alpha,t}$ LiFi user connections and WiFi user connections $\mu_{\beta,t}$, and the bandwidth proportioned by the LiFi APs and WiFi AP, represented by $b_{\mu,\alpha,t}$ and $b_{\mu,\beta,t}$ respectively. The RWP model used a constant velocity for all user mobility, and the users never hesitated and were constantly moving, similar to the *constant speed* RWP model[2]. The general characteristics of the environment can be seen in Table 3.1 and illustrated in Figure 3.1.

3.2 Optimization Function

This paper is based on data rate calculations on the LoS path in the LiFi channel model. The influence of reflected paths, that being constructive or destructive, was not included when calculating user data rates, however, interference from

Parameter	Value
Room Dimensions	5mx5mx3m
Number of Users	{12,20}
LiFi APs	4
LiFi Bandwidth	20 MHz
LiFi Max Rate	250 Mbps
WiFi APs	1
WiFi Bandwidth	20 MHz
WiFi Max Rate	160 Mbps

Table 3.1: Environment Values

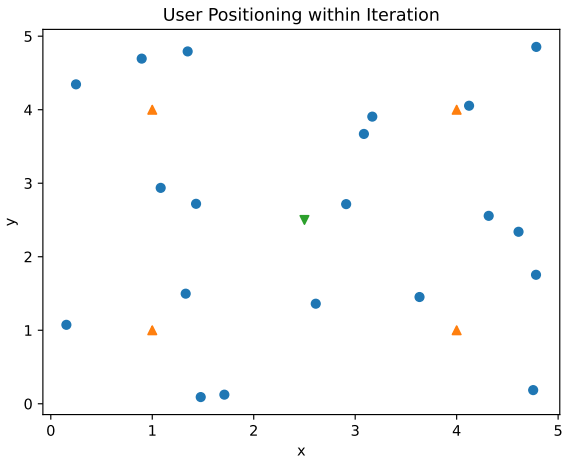

Figure 3.1: Environment during an iteration, with 20 users •, a centralized WiFi AP ▽, and 4 spread LiFi APs △

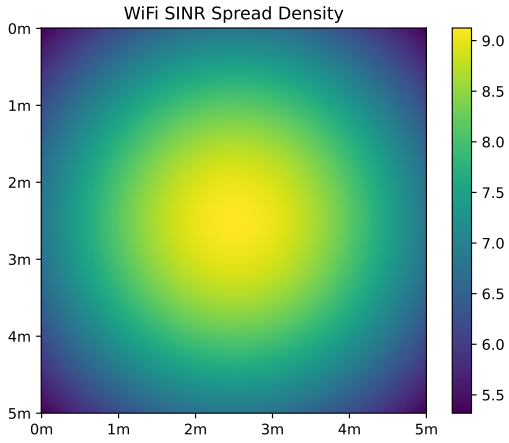

Figure 3.2: This demonstrates the spread of the WiFi SNR available to users at a constant height of $1,4$m

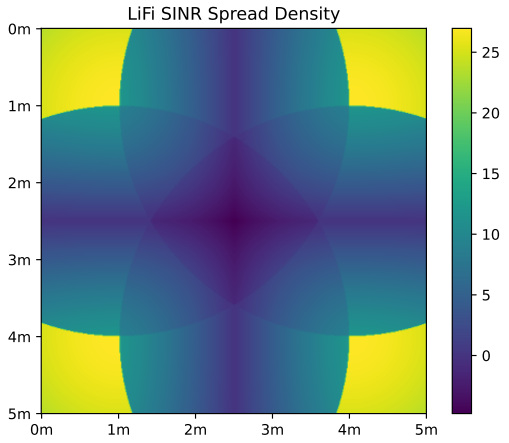

Figure 3.3: This demonstrates the spread of the maximum LiFi SINR available to users at a constant height of $1,4$m

Parameter	**Value**
Refractive Index χ	1.5
Gain of Optical Filter T_s	1.0
Area of PD	100 mm^2
Concentrator Gain $g(\psi)$	1.0
Optical Power P_{opt}	5 W
Max Data Rate $r_{LiFi,max}$	250 Mbps
Noise Power N_{LiFi}	10^{-21} A^2/Hz

Table 3.2: Parameters of the LiFi channel calculations

other signals was taken into account; the path loss calculations were done based on a *Free-Space Path Loss Model*. The LiFi channel path loss was estimated as follows[4]:

$$H^t_{LiFi(\mu,\alpha)} = \begin{cases} \frac{(m+1)A}{2\pi d^2} cos^m(\phi) T_s(\psi) g(\psi) cos(\psi), & 0 \leq \psi \leq \Psi_f \\ 0, & \psi \geq \Psi_f. \end{cases} \tag{3.1}$$

here m represents the Lambertian order 3.2, from the function of the half-intensity radiation semi-angle $\phi_{1/2}$. Which refers to the angular range of the signal, within which the intensity of the signal wave emitted is reduced by half of its maximum value along a specific direction. The physical area of the PD is represented by A and the distance between the LiFi AP and the PD is d. In the code d is calculated from the Eucledian norm of a distance vector between the users and an AP. The ψ and μ represent the angle of the signal against the PD. And lastly, Ψ_f represents the receiver Field of View (FoV), beyond which, there is no LoS path and no signal transmission. The $T_s(\psi)$ and $g(\psi)$ are the gain of the optical filter and the gain of the concentrator respectively, both of which are negligible.

$$m = \frac{-1}{log_2(cos(\phi_{1/2}))} \tag{3.2}$$

Due to a WiFi signal being more adept to a Non-Line-of-Sight (NLoS) scenario - in particular, due to larger wavelengths, allowing for greater diffraction - the angle of incidence between the PD and the signal did not need to be considered in the path loss calculations. Within a *Free-Space Loss Model*, a WiFi signal is primarily influenced by the distance of the UD to the AP. The path loss model for the WiFi channel in decibels was calculated as follows:

Parameter	Value
Shadow Fading X_{sf}	3.0
Radio Frequency f_{rf}	2.45 GHz
Distance Breakingpoint d_{bp}	10 m
Max Data Rate $r_{WiFi,max}$	160 Mbps
Transmission Power P_T	0.1 W
Noise Power N_{WiFi}	10^{-15} A^2/Hz

Table 3.3: Parameters of the WiFi channel calculations

$$H^t_{WiFi(\mu,\beta)} = \begin{cases} 20\log_{10}(d) + 20\log_{10}(f_{rf}) - 20\log_{10}(\frac{4\pi}{c}) + X_{sf}, & d < d_{bp} \\ 20\log_{10}(d_{bp}) + 20\log_{10}(f_{rf}) - 20\log_{10}(\frac{4\pi}{c}) + 35\log_{10}\left(\frac{d}{d_{bp}}\right) + X_{sf}, & d \geq d_{bp} \end{cases} \tag{3.3}$$

Where X_{sf} represents the shadow fading, which is the random variation in the received singal due to local obstacles and environmental conditions. Shadowing is a form of slow large-scale fading, and alters signal strength. f_{rf} represents the radio frequency of the WiFi signal. In this equation, it is assumed that the signal energy is spread over a spherical AP, but in the calculations done by the model an approximation of 147.5 for the free space path loss at a frequency of 1 GHz. Beyond the d_{bp} breakpoint distance, the WiFi channel calculations begin to take into account the diffraction and scattering of the signal, and the increase path loss caused by these factors.

With the path loss of the WiFi and LiFi channels, the signal-to-noise ratio (SNR) of both models can be calculated. The LiFi SNR calculation is computed as follows[4][10][11]:

$$SNR^t_{LiFi(\mu,\alpha)} = \frac{(H^t_{LiFi(\mu,\alpha)} P_{opt} \kappa)^2}{N_{LiFi} B_{LiFi}} \tag{3.4}$$

where P_{opt} denotes the transmitted optical power, and κ is the system's optical-to-electrical conversion efficiency, N_{LiFi} is the noise power spectral density (NSPD) at a LiFi frequency, and B_{LiFi} is the bandwidth of the LiFi AP α. It is relevant to take into account however, the potential interference from signals of other LiFi APs in the environment, all of which function within similar frequencies. Therefore it is important to consider the LiFi signal-to-interference-noise ratio (SINR) 3.5,

$$SINR^t_{LiFi(\mu,\alpha)} = \frac{(H^t_{LiFi(\mu,\alpha)} P_{opt} \kappa)^2}{N_{LiFi} B_{LiFi} + \sum_{\gamma \epsilon \{1,..,4\}/\{\alpha\}} ((H^t_{LiFi(\mu,\alpha)} P_{opt} \kappa)^2)} \tag{3.5}$$

which includes the channel gain of the interfering LiFi APs γ with the user μ. With the SINR of the LiFi channel, the maximum achievable data rate for the user from the LiFi channel can be calculated with the following:

$$r_{LiFi(\mu,\alpha)} = \begin{cases} \frac{B}{2}\log_2(1 + (\frac{e}{2\pi}SNR^t_{LiFi(\mu,\alpha)}), & r_{LiFi(\mu,\alpha)} < r_{LiFi,max} \\ r_{max}, & r_{LiFi(\mu,\alpha)} \geq r_{max} \end{cases} \tag{3.6}$$

Regarding the SNR calculation for the WiFi channel, it is given by the following equation:

$$SNR^t_{LiFi(\mu,\beta)} = \frac{(H^t_{WiFi(\mu,\beta)})^2 P_T}{N_{WiFi} B_{WiFi}} \tag{3.7}$$

where N_{LiFi} denotes the WiFi NSPD, and B_{WiFi} is the bandwidth of the AP α. As there are no additional WiFi APs within the environment, there is no interference from other channels, as *Multipathing* is not being considered. This allows a direct calculation of the user data rate, using the *Shannon Capacity*:

$$r_{WiFi(\mu,\beta)} = \begin{cases} B\log_2(1 + (SNR^t_{WiFi(\mu,\beta)}), & r_{WiFi(\mu,\beta)} < r_{WiFi,max} \\ r_{WiFi,max}, & r_{WiFi(\mu,\beta)} \geq r_{WiFi,max} \end{cases} \tag{3.8}$$

The optimization function is derived from the accumulative data rate for every user, over all time iterations. To compute the data rates over a single-time service, the connections decided by the CU must be taken into consideration. As we are implementing a heterogeneous HHO network, users are capable of connecting to a WiFi AP and LiFi AP during the same service time, so the data rates from the LiFi channel in equation 3.9 and WiFi channel in equation 3.9 need to be separately determined.

$$R^t_{LiFi} = \sum_{\alpha \epsilon APs, \mu \epsilon users} x^t_{\mu,\alpha} r_{LiFi(\mu,\alpha)} \eta^t_{\mu,\alpha} b_{LiFi,\mu} \tag{3.9}$$

$$R^t_{WiFi} = \sum_{\beta \epsilon APs, \mu \epsilon users} y^t_{\mu,\beta} r_{LiFi(\mu,\beta)} \eta^t_{\mu,\beta} b_{WiFi,\mu} \tag{3.10}$$

$x^t_{\mu,\alpha}$ and $y^t_{\mu,\beta}$ can be 1 or 0, and are meant to represent the decision of the CU, whether to connect a user μ to AP $\alpha|\beta$ at the time service t. Additionally, $\eta^t_{\mu,\alpha|\beta}$ represents whether a handover should occur at this iteration or not, determined by equation 3.13. A handover multiple of 0.9 is used [3], as handovers have less detrimental effects when occurring within the same technology. Handovers occur only if previously the user μ was not already connected to this AP $\alpha|\beta$; handovers occur predominantly within the LiFi channel, due to the limited range capacity of

the LiFi APs. Lastly, $b_{(LiFi|WiFi),\mu}$ represents the bandwidth proportioned to the user by the CU.

$$x^t_{\mu,\alpha_0} = 1 \Rightarrow x^t_{\mu,\alpha} = 0, \qquad \forall \alpha \epsilon APs, \alpha \neq \alpha_0 \tag{3.11}$$

$$y^t_{\mu,\beta} = 1 \Rightarrow y^t_{\mu,\beta} = 0, \qquad \forall \beta \epsilon APs, \beta \neq \beta \tag{3.12}$$

$$\eta^{t+1}_{\mu,\alpha} = \begin{cases} 0.9, & \alpha^{t+1}_{\mu} = \alpha^t_{\mu} \\ 1, & otherwise \end{cases} \tag{3.13}$$

To conclude, the final reward received by the RL model would be the accumulated data rate of both channels over the specified period. This makes the model prioritize the overall user experience rather than the instantaneous data rate.

$$R_{accum} = \sum_{t \epsilon T} R^t_{WiFi} + R^t_{LiFi} \tag{3.14}$$

3.3 Action and Observation Space

The observation spaces of all models are similarly constructed, being solely dependent on the number of users and APs within the environment, and whether a proactive approach was applied. The model would be informed of the SINR associated to each user μ and every α LiFi AP and β WiFi AP.

$$\begin{bmatrix} SINR^{t_0}_{\mu,\alpha} & ... & ... \\ ... & ... & ... \\ ... & ... & ... \\ SINR^{t_0}_{\mu,\beta} & ... & ... \end{bmatrix} \mu \epsilon \{1, ..., \mu_{max}\}, t \epsilon \mathbb{N}, \alpha \epsilon \{1, ..., 4\}, \beta \epsilon \{1\}; \tag{3.15}$$

Where t represents the time iteration of the model, and α and β the Lifi and WiFi APs respectively. In a proactive approach, the model is considered to have a predictive capability of the user positions. To simulate this, the observation space is extended by a window size w_pro of 5 iterations. This allows the model to avoid potential overwhelming demand on a single AP, by allocating resources preemptively, and the model is expected to learn how to prevent excessive handovers. In this case, the observation space would have 5 additional matrices 3.16.

$$\begin{bmatrix} SINR_{\mu,\alpha}^{t_0} & ... & ... \\ ... & ... & ... \\ ... & ... & ... \\ SINR_{\mu,\alpha}^{t_0+w_{pro}} & ... & ... \\ ... & ... & ... \\ ... & ... & ... \\ SINR_{\mu,\beta}^{t_0} & ... & ... \\ ... & ... & ... \\ SINR_{\mu,\beta}^{t_0+w_{pro}} & ... & ... \end{bmatrix} \mu\epsilon\{1, ..., \mu_{max}\}, t\epsilon\mathbb{N}, \alpha\epsilon\{0, ..., 4\}, \beta\epsilon\{0, 1\}; \quad (3.16)$$

Regarding the action space, it differs for *Multidiscrete*, *Binary*, and *Continuous* RL models:

1. *Continuous*: with a continuous action space, the CU had the most command over the resource allocation. This made it more difficult for the models to learn, as the computational cost and complexity were the highest. The bandwidth, in this case, could be preserved, as only the amount specified by the model was used, although if the bandwidth demand, set by the CU, for an AP, was greater than 100, all bandwidth values associated with this AP were set to 0.

2. *Multidiscrete*: in this method, the model was capable of connecting any user to any AP at all iterations; however, in order to simplify the learning process, the bandwidth distribution was done in a proportionate manner. This means that all available bandwidth was used, and distributed according to values proportioned by the CU. This prevented the model from being penalized when demanding an accumulated bandwidth over 100 from an AP and allowed it to prioritize user packages with more bandwidth.

3. *Binary*: this is the simplest action space used, and had the lowest computational cost of all the RL models. The *binary* method also used a *MultiDiscrete* action space, but the connection values were limited to binary. This was to limit the CU's options when deciding which AP to connect a user to. In this method, it was only capable of deciding whether or not to connect the user to the AP - from the LiFi and WiFi channels - with the highest SINR at that iteration.

$$\begin{bmatrix} \alpha_{\mu}^{t_0} & ... & ... \\ \beta_{\mu}^{t_0} & ... & ... \\ b_{LiFi,\mu}^{t_0} & ... & ... \\ b_{WiFi,\mu}^{t_0} & ... & ... \end{bmatrix} \mu\epsilon\{1, ..., \mu_{max}\}, t\epsilon\mathbb{N}, \alpha\epsilon\{0, ..., 4\}, \beta\epsilon\{0, 1\}, b\epsilon\{0, ..., 100\}; \quad (3.17)$$

3.4 Training

Various RL models were created and analyzed during the research of this paper. These models vary in proactive and reactive approaches, different RL algorithms, and the adjustments applied to the action space. The environmental setup for these models was done in the *OpenAi Gym* and *Unity ML-Agents* python libraries. This provided a structured standardized environment, made up of a *reset*, *initialization* and *step function*. A realistic simulation of 4 LiFi APs and 1 WiFi AP providing packages to users was established, by every new iteration rerunning the *step function*. After the building of the environment, the various models went through hyperparameter tuning; the key hyperparameters established were (*learning rate*, *entropy coefficient*, *discount factor*.

1. *learning rate*: It controls how much the agent adjusts its policy and value function estimates in response to the rewards of the simulation. This is particularly important to reduce computational costs and to achieve the true ideal convergence.

2. *entropy coefficient*: Entropy coefficient balances exploration and exploitation by encouraging the policy to be more exploratory; this, in turn, reduces the stability of the results.

3. *discount factor*: The discount factor determines the importance of future rewards in the agent's decision-making process.

The hyperparameter tuning is done using the user positions from the *evaluation* scenario, which consists of 500 iterations. After the best hyperparameters are defined, the model undergoes training, which has its own set of user positions, being made up of 1500 iterations. The training ends once convergence is reached, seen in figure 3.4.

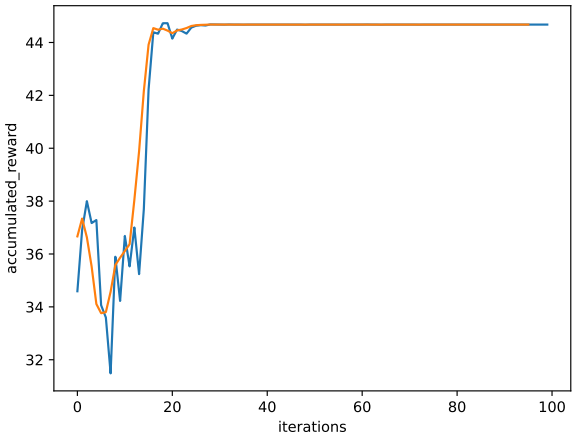

Figure 3.4: Model training example, where the training stops once the accumulated reward for each simulation reaches a convergence

Chapter 4

Performance Evaluation

In this paper, various RL models were analyzed, in order to conclude in which situations an RL model is suited for resource allocation. The ideal RL model was discovered through testing various adaptions; the key characteristics of the model that were altered were the action space, the algorithm, and proactivity. These models had varying computational costs, and also very diverse sets of results.

4.1 Varying Action Spaces

The action space refers to the set of possible actions that the agent can take in its environment. As discussed in Chapter 3, the action space values define the AP and user connections, as well as the bandwidth proportioned to each user. A larger and more diverse action space allows the CU to have a more expressive command over resource allocation, which can be important for handling complex tasks. On the other hand, a complex action space can make learning more challenging due to the increased number of choices. The action space is extremely influential on the policies learned by the model and can be decisive in the model reaching more desirable results.

4.1.1 Continuous

The continuous model had the highest computational cost during training, due to the complexity of the model's action space. In return, this model had the most control over the resource allocation of every simulation. Due to these exorbitant computational costs, the training never reached convergence and was very unstable. This method promotes a more exploratory training method, nevertheless, even after additional training simulations, the model could not learn an ideal method

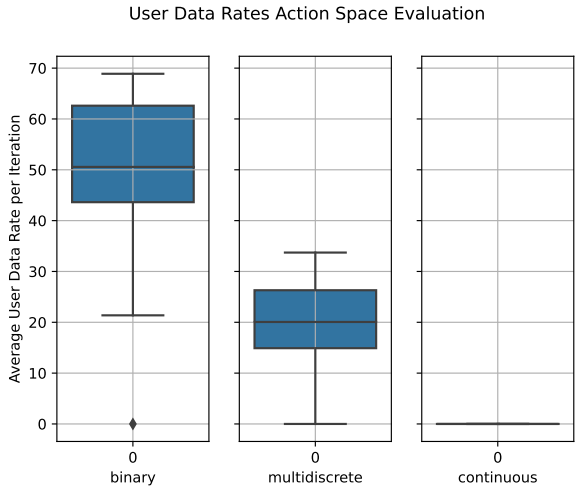

Figure 4.1: The spread of average user data rate per iteration

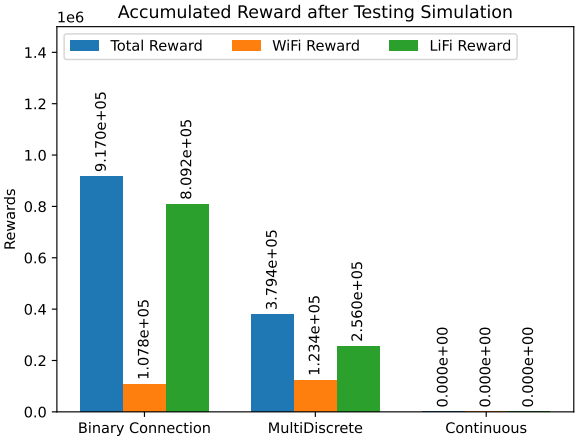

Figure 4.2: The accumulated reward for the different action spaces

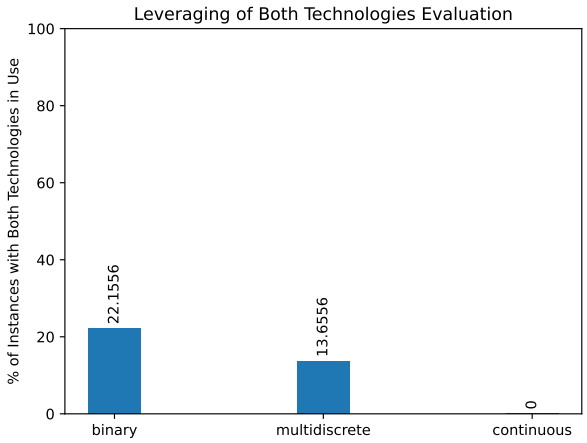

Figure 4.3: The total number of instances of users connected to both technologies for the different action spaces

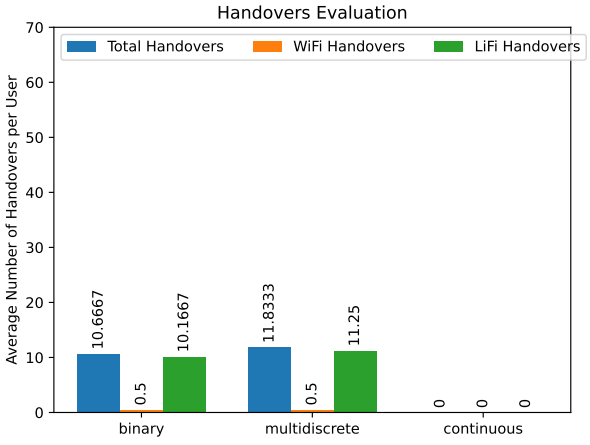

Figure 4.4: Total number of handovers for the different action spaces

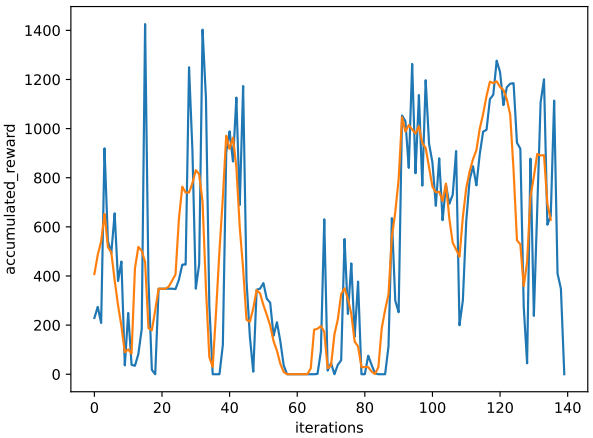

Figure 4.5: Training progression for proactive continuous model

to distribute bandwidth and establish wireless connections, Figure 4.5. Even after 140 iterations, the model was unable to reach a convergence, and, when introduced to the testing environment, this model had no success.

4.1.2 MultiDiscrete

The multidiscrete model was the most complex RL model among the different action spaces used, that was able to reach a convergence during training, seen in Figure 4.6. This model also applied the *proportional bandwidth* method, previously explained in Chapter 3. This simplification method made it viable for the model to learn how to distribute bandwidth among users connected to the same AP. Regarding the results in the testing phase, the model had an accumulated data rate of $380,000$ *bps*. The model seemed to be very unstable, and although having reached a convergence, there was still a clear difficulty in connecting to the correct APs. This is seen by the excessive number of handovers in comparison to the binary connection model in Figure 4.4. In various iterations, although there were viable SINR values in the observation space, the model left LiFi APs unemployed. For clarification, Table 4.1 shows the connections established with the corresponding observation space 5 users within the 250^{th} service time. There are numerous examples of the model connecting to the wrong AP, and not using all available resources.

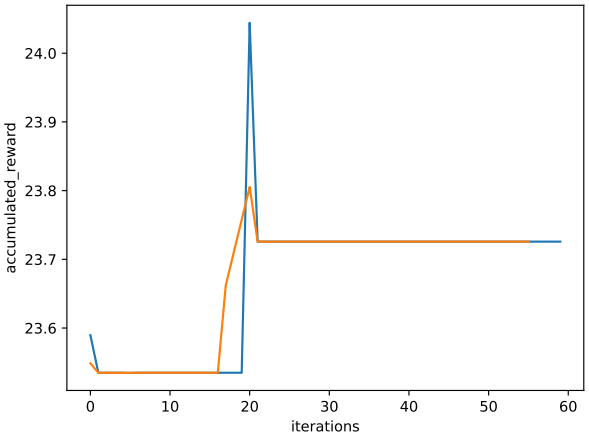

Figure 4.6: Training progression for proactive multidiscrete model

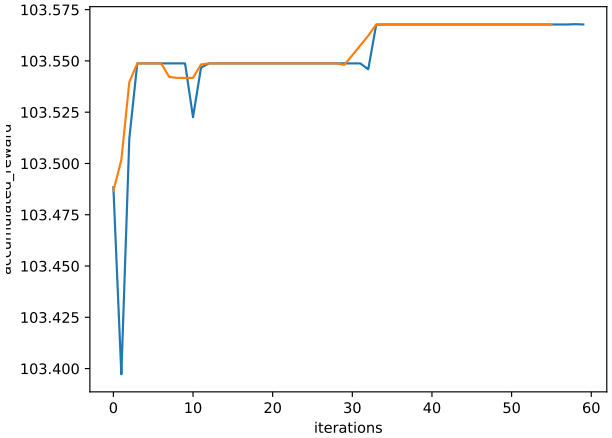

Figure 4.7: Training progression for binary connection model

LiFi Connections	LiFi SINR
2	[0, 33.14, 0, 0]
3	[0, 32.52, 0, 0]
1	[0, 33.01, 0, 0]
3	[35.01, 0, 0, 0]
2	[0, 0, 0, 1.34]
3	[0, 0, 0, 37.2]

Table 4.1: Extensive analysis of the service time 250 for the MultiDiscrete model; LiFi SINR is the observation space matrix, so the connection n will use the n^{th} value in the matrix to calculate user data rate

4.1.3 Binary Connection

In order to resolve this issue of wasting unused resources, a binary action space was implemented. In this model, the CU could only decide whether or not to connect to the nearest AP to the user. For our environment, this was not too limiting, as users rarely had the potential for multiple connections within the LiFi channels. The bandwidth allocation remained the same as in the multidiscrete action space, using the proportional bandwidth method. In training, this model had the least computational cost among the various action spaces, and reached a convergence very early during training, roughly at the 30th iteration as seen in Figure 4.7. Finally, within the testing phase, there were clear improvements in the decision-making of the model. The accumulated data rate was nearly 3 times as large as that of the multidiscrete model; this, however, was not clear if it was due to a more intelligent model, or due to the lower probability of leaving APs unused, due to the limited action space. It is also noticeable this model has fewer instances of idle APs from Figure 4.3, where the simultaneous use of both LiFi and WiFi channels is considerably larger than the multidiscrete model.

4.2 Reinforcement Learning Algorithms

The choice of algorithm can have a profound impact on how an algorithm learns and adapts to the environment. In this paper, algorithms with dissimilar policies were tested, to have a broad overview of the effects of the algorithm on the results of the model. The key differentiating characteristics included the exploration policy, the algorithm stability, and policy-based and value-based methods.

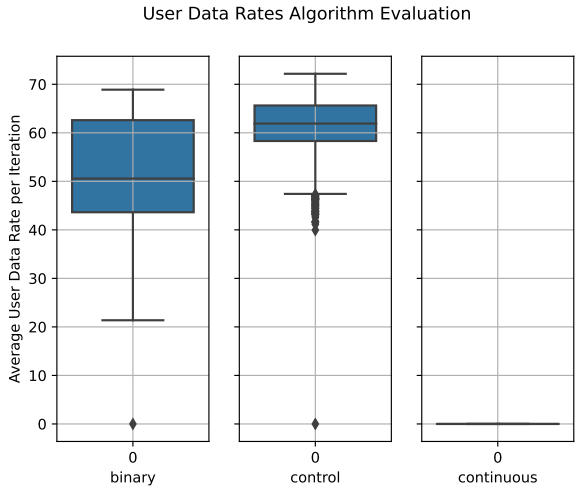

Figure 4.8: The spread of average user data rate per iteration

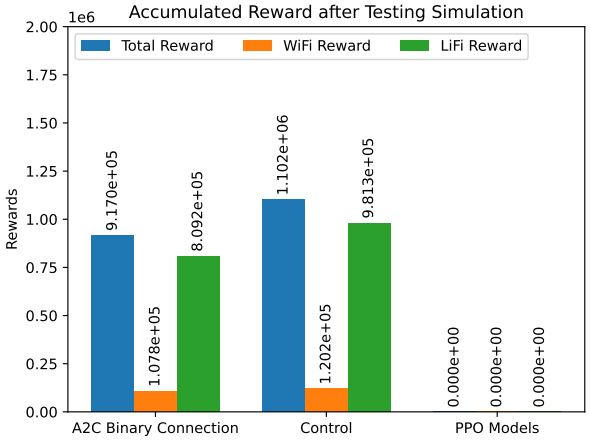

Figure 4.9: The accumulated reward for the different algorithms

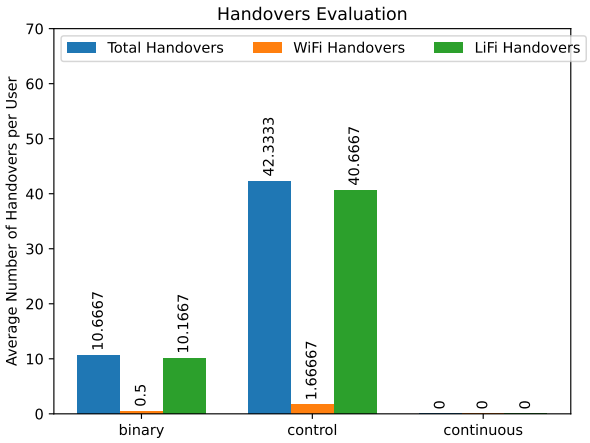

Figure 4.10: Total number of handovers for the different algorithms

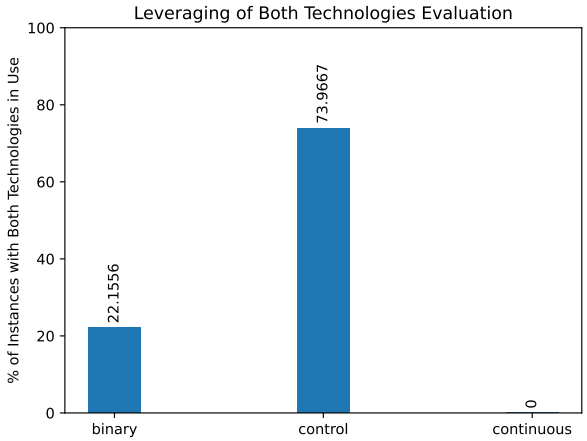

Figure 4.11: The total number of instances of users connected to both technologies for the different action spaces

4.2.1 A2C

A2C combines policy-based and value-based approaches. This makes it an ideal algorithm to test various action spaces on, both continuous and discrete. The continuous A2C model, as discussed previously, was not able to learn even with the extended training period provided, as a consequence of the complexity of the model. As seen in Figure 4.9 the binary model, which uses a discrete action space, had more success, although it had poorer results than the logic-based control model.

4.2.2 PPO

In contrast, PPO is purely a policy-based algorithm. This means it will try to learn the correct policy to find the ideal values, particularly useful in continuous action spaces. For the continuous example, the same issues in training were still prevalent, even with an extended training period, the model could not reach convergence, evident in Figure 4.13. When simulated in the testing environment, it had very poor results, similar to the continuous A2C model. Regarding the PPO multidiscrete model, although training neared a convergence, unlike the continuous model, it was still very unstable; this was probably due to the PPO algorithm's policy-based nature. In testing, the model provided no results.

4.2.3 Control

To have a better assessment of how well the reinforcement learning models were distributing resources, a control environment was created, based purely on a formulated logic. The environment would connect to the LiFi and WiFi AP that provided the largest SINR value, and the bandwidth would be equally spread among all connected users. The final accumulated user data rate was higher than the best reinforcement learning model, the proactive A2C binary connection model. These results do not justify using an RL model for resource allocation in an HLWN. As seen in Figure 4.11, the control model had many more instances with simultaneous use of both LiFi and WiFi channels, however, the relatively minimal difference accumulated may suggest that the RL model allocates bandwidth more appropriately. Additionally, a handover multiple of 0.9 was used, but if the technologies were less efficient in handovers the accumulated reward of the control model might drop more drastically as it had more handovers, evident in Figure 4.10.

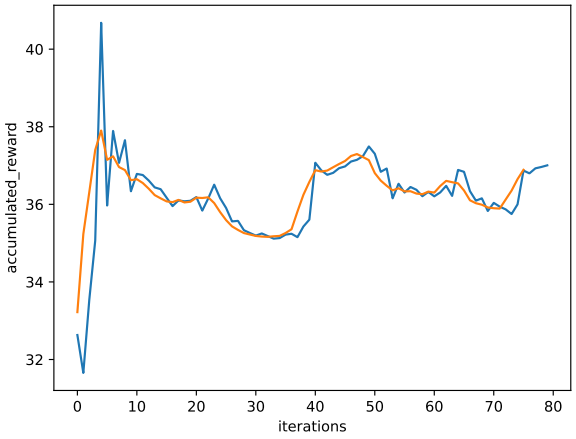

Figure 4.12: Training progression for PPO MultiDiscrete model

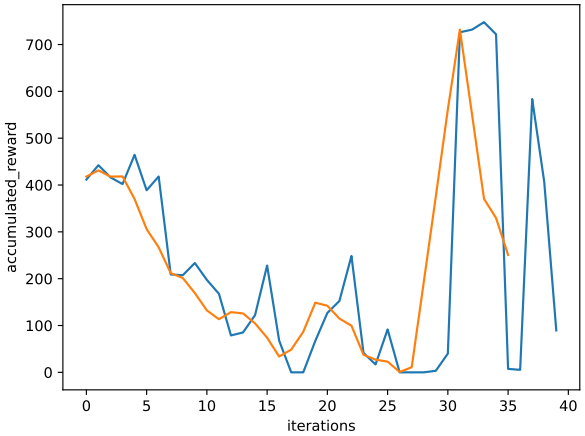

Figure 4.13: Training progression for PPO continuous model

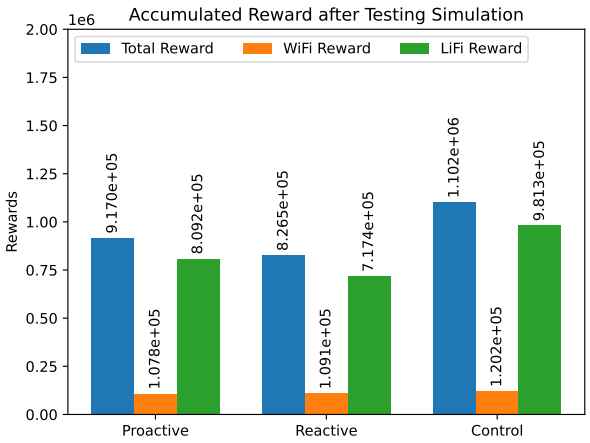

Figure 4.14: The accumulated reward for the different algorithms

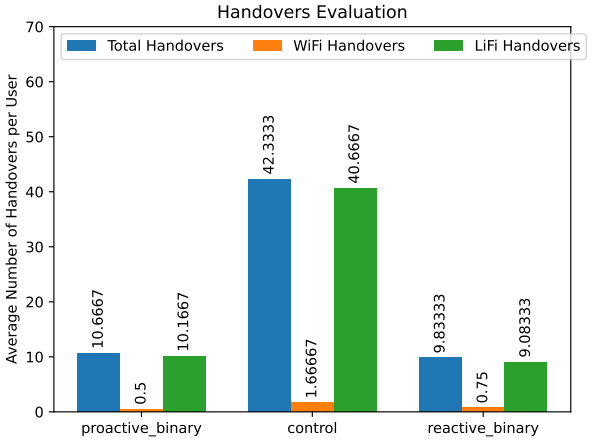

Figure 4.15: Total number of handovers for the different algorithms

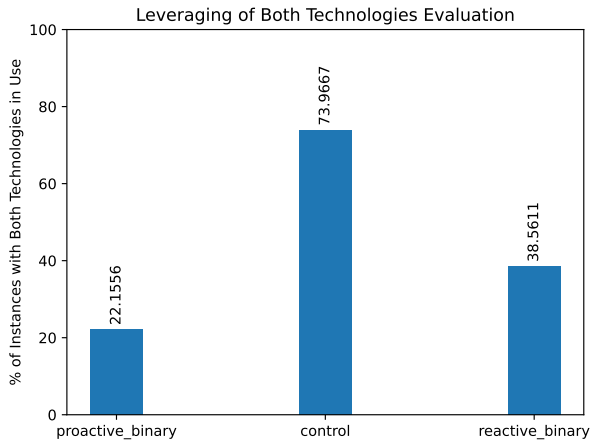

Figure 4.16: The total number of instances of users connected to both technologies for the different action spaces

4.3 Proactivity

The proactivity of the model is expected to greatly influence the number of handovers occurring during the simulation. It is expected to be able to avoid overexertion of APs, and minimize the number of handovers, by preemptively allocating packages to neighboring APs.

4.3.1 Proactive

In the proactive examples mentioned previously, the model with the most promising results was the binary connection model, with an A2C algorithm. This model was the most successful in distributing user demand within the network, primarily by not leaving an AP idle when an SINR value is detected as often. The proactivity of the model is potentially very important, to preemptively allocate users to different APs to preserve the bandwidth for future iterations. There was a clear improvement in the total data rates of the proactive models compared to the reactive models, although the reactive models had more instances connected to both technologies. This is due to the avoidance of overexertion of APs by the proactive model, to allow higher data rates in future iterations.

4.3.2 Reactive

An identical model, with binary connection and an A2C algorithm, was created with a reactive observation space, so no window of future user positions was avail-

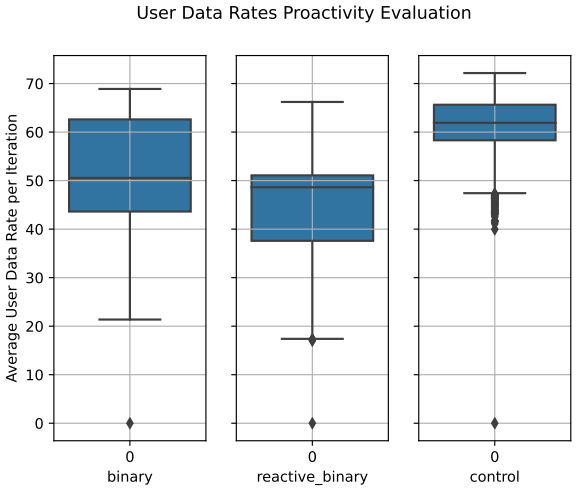

Figure 4.17: The spread of average user data rate per iteration

able to the model. This in theory wouldn't allow the model to knowingly prevent future blockages within the system. The computational cost however was extremely minimal, as the model reached a convergence within the first initial iterations. As seen in 4.14, the reward progression is slightly worse than the proactive model as expected. This roughly 10% increase in data rate may justify the use of a proactive approach in RL models for resource allocation, however, it also has to be taken into account, that a predictive model would have a margin of error which would influence the proactive model results.

Chapter 5

Conclusions and Outlook

In conclusion, this research endeavor embarked on the creation and evaluation of an RL model for the proactive allocation of resources within an HLWN. The objective was to develop an intelligent system that optimally manages resource allocation to enhance network performance and user satisfaction. The significance of this work stems from the increasing demand for efficient utilization of network resources in modern communication systems, and the lack of bandwidth in currently used RF networks.

The process commenced with developing various RL models, each tailored to tackle the resource allocation challenge. After extensive training and testing, it became evident that the binary connection model exhibited the most promising results among the RL models. Remarkably, this model not only demonstrated impressive performance but also exhibited the advantage of having low computational costs, offering a viable solution that aligns with the real-time demands of network management. Despite its computational simplicity, the binary connection model had marginal performance improvement over a basic logic-based resource allocation scheme.

Among the reinforcement learning algorithms tested, A2C emerged as the most effective choice. The broad nature of A2C, incorporating both value-based and policy-based elements, allowed for a more stable training process resulting in models more suited to resource allocation in an HLWN. In contrast, the PPO algorithm exhibited certain limitations, possibly stemming from the purely policy-based methods it employs, highlighting the necessity for algorithm selection, which demands a high computational cost in training, as it invokes a more exploratory method.

In the future, this research can guide others in making resource management sys-

tems better. As technology keeps improving, the lessons from this study can help mix machine learning and logic to ensure HLWN use its resources wisely, adapting to dynamic changes smoothly.

Appendix A

Notations and Abbreviations

AP	Access Point
CU	Control Unit
FL	Fuzzy Logic
FoV	Field of View
HHO	Horizontal Handover
LoS	Line of Sight
MANET	Mobile Ad Hoc Network
NLoS	Non-Line-of-Sight
PCR	Predictive Channel Reservation
PD	Photo Detector
RL	Reinforcement Learning
RWP	Random Waypoint
SINR	Sounding Reference Signal
SNR	Signal-to-Noise Ratio
TC	Topology Control
TD	Threshold Distance
TORA	Temporally-Ordered Routing Algorithm
UD	User Device
VHO	Vertical Handover
ZRP	Zone Routing Protocol

Bibliography

[1] Harald Haas. Lifi is a paradigm-shifting 5g technology. *Reviews in Physics*, 3:26–31, 2018.

[2] Xiping Wu and Harald Haas. Load balancing for hybrid lifi and wifi networks: To tackle user mobility and light-path blockage. *IEEE Transactions on Communications*, 68(3):1675–1683, 2019.

[3] Jaafaru Sanusi, Sadiq Idris, Abiodun Musa Aibinu, Steve Adeshina, and Ali Nyangwarimam Obadiah. Handover in hybrid lifi and wifi networks. In *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–6. IEEE, 2019.

[4] Mohammad Amir Dastgheib, Hamzeh Beyranvand, Jawad A Salehi, and Martin Maier. Mobility-aware resource allocation in vlc networks using t-step look-ahead policy. *Journal of Lightwave Technology*, 36(23):5358–5370, 2018.

[5] Sunil Kumar Singh, Rajesh Duvvuru, and Amit Bhattcharjee. Performance evaluation of proactive, reactive and hybrid routing protocols with mobility model in manets. *International Journal of Research in Engineering and Technology (IJRET)*, 2(8):254–259, 2013.

[6] Michael J Neely. Universal scheduling for networks with arbitrary traffic, channels, and mobility. In *49th IEEE Conference on Decision and Control (CDC)*, pages 1822–1829. IEEE, 2010.

[7] Ming-Hsing Chiu and Mostafa A Bassiouni. Predictive schemes for handoff prioritization in cellular networks based on mobile positioning. *IEEE Journal on selected areas in communications*, 18(3):510–522, 2000.

[8] Liqiang Wang, Dahai Han, Min Zhang, Danshi Wang, and Zhiguo Zhang. Deep reinforcement learning-based adaptive handover mechanism for vlc in a hybrid 6g network architecture. *IEEE Access*, 9:87241–87250, 2021.

[9] Xiping Wu, Majid Safari, and Harald Haas. Three-state fuzzy logic method on resource allocation for small cell networks. In *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1168–1172. IEEE, 2015.

[10] Yunlu Wang, Dushyantha A Basnayaka, Xiping Wu, and Harald Haas. Optimization of load balancing in hybrid lifi/rf networks. *IEEE Transactions on Communications*, 65(4):1708–1720, 2017.

[11] Rizwana Ahmad, Mohammad Dehghani Soltani, Majid Safari, Anand Srivastava, and Abir Das. Reinforcement learning based load balancing for hybrid lifi wifi networks. *IEEE Access*, 8:132273–132284, 2020.