Received 4 May 2022; revised 24 May 2022; accepted 9 June 2022. Date of publication 14 June 2022; date of current version 27 June 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3183172

Optimizing Packet Forwarding Performance in Multiband Relay Networks via Customized Reinforcement Learning

BUSHRA MUGHAL [1] (Graduate Student Member, IEEE), ZUBAIR MD. FADLULLAH [2,3] (Senior Member, IEEE), MOSTAFA M. FOUDA [4] (Senior Member, IEEE), AND SALAMA IKKI [1] (Senior Member, IEEE)

[1]Department of Electrical Engineering, Lakehead University, Thunder Bay, ON P7B 5E1, Canada

[2]Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada

[3]Thunder Bay Regional Health Research Institute (TBRHRI), Thunder Bay, ON P7B 7A5, Canada

[4]Department of Electrical and Computer Engineering, College of Science and Engineering, Idaho State University, Pocatello, ID 83209, USA

CORRESPONDING AUTHOR: B. MUGHAL (e-mail: bmughal@lakeheadu.ca)

ABSTRACT In next-generation wireless networks, relay-based packet forwarding, emerged as an appealing technique to extend network coverage while maintaining the required service quality. The incorporation of multiple frequency bands, ranging from MHz/GHz to THz frequencies, and their opportunistic and/or simultaneous exploitation by relay nodes can significantly improve system capacity, however at the risk of increased packet latency. Since a relay node can use different bands to send and receive packets, there is a pressing need to design an efficient channel allocation algorithm without a central oracle. While existing greedy heuristics and game-theoretic techniques, which were developed for multi-band channel assignment to relay nodes, achieve minimum packet latency, their performance drops significantly when network dynamism (i.e., user mobility, non-quasi-static channel conditions) is introduced. Since this problem involves multiple relay nodes, we model it as a Markov Decision Process (MDP) involving various stages, which essentially means that achieving an optimal and stable solution is a computationally hard problem. Since solving the MDP, traditionally, consumes a great deal of time and is intractable for relay nodes, we explore how to approximate the optimal solution in a distributed manner by reformulating a reinforcement learning-based, smart channel adaptation problem in the considered multi-band relay network. By customizing a Q-Learning algorithm that adopts an epsilon-greedy policy, we can solve this re-formulated reinforcement learning problem. Extensive computer-based simulation results demonstrate that the proposed reinforcement learning algorithm outperforms the existing methods in terms of transmission time, buffer overflow, and effective throughput. We also provide the convergence analysis of the proposed model by systematically finding and setting the appropriate parameters.

INDEX TERMS Machine learning, reinforcement learning, Markov decision process, Q-learning, multiband communication, relay network.

I. INTRODUCTION

IN MODERN wireless networks, relay communication is an extensively employed and researched network topology [1]. The main motivation behind using a relay network is to extend the coverage area and improve the service outage probability. In scenarios where a source node (SN) needs to transmit certain packets to a destination node (DN) and experiences a weak wireless link (due to high interference and poor channel conditions), a relay node (RN) can be employed as an intermediate node to forward the data packet to the DN. The idea of a sequence of RNs can be used to widen the network coverage in cases where the SN and DN are

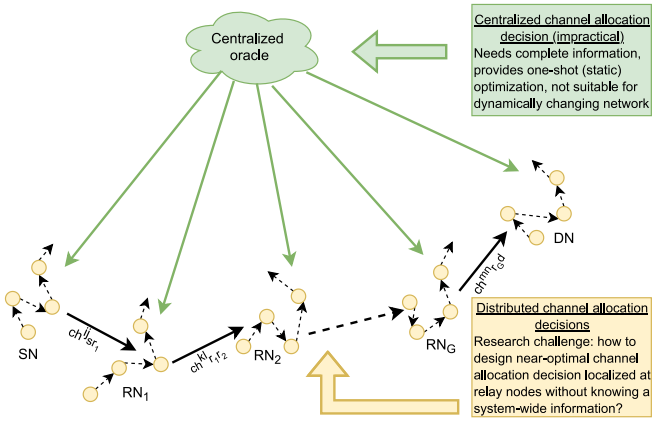


FIGURE 1. Centralized vs distributed channel allocation in mobile multi-band relay network.

completely out of communication. However, in a conventional relay network, using an intermediate node typically introduces a higher latency to the SN-DN packets. The reason for this is that the RN first waits to completely receive data packets from SN, and once they are fully received, the same dedicated channel is now used by the RN to forward them to the DN. This conventional method of relaying data is referred to as Decode and Forward (DF) [2].

To mitigate the shortcomings of the conventional relay topology, instead of waiting to fully receive the packets from the SN, and then waiting for the dedicated channel to become unoccupied, and then forwarding the same data packet to its DN, the RN can immediately start forwarding data on a different channel of any other frequency band for simultaneous reception and transmission. This phenomena can be seen in Fig. 1, where the RN receives data from the SN at the i^{th} channel of the j^{th} band ($\text{ch}_{sr_1}^{ij}$) and forwards the same data to the next node via the k^{th} channel from a different l^{th} band ($\text{ch}_{r_1r_2}^{kl}$).

This is considered multi-band communication. With modern technology, we can assume the presence of radios capable of supporting a variety of transceiving frequency bands (e.g., IEEE 802.11ax operating at 2.4GHz and 5GHz [3], millimeter wave (mmWave), visible light communication, and so forth) that react differently to path loss, fading, mobile blocking, and other physical phenomena. As a consequence, there is a large diversity among channel conditions across different frequency bands [4], [5]. The utilization of multi-band communication requires efficient channel access and buffer management techniques/approaches to handle the differences in data rate when receiving and forwarding packets at the same time. We also need to consider how to access the scarce spectrum accordingly. This results in a computationally hard problem, where we need to optimize channel allocation using the available multi-bands to minimize the data packet latency in the relay network.

Since multi-band channel allocation to RNs is a time-sensitive problem, a suitable (preferably optimal) decision needs to be made without introducing more delay to the problem-solving process. As seen in Fig. 1, a decision needs to be made with regards to the frequency channels used by the RNs to forward the data to the DN. If considered centrally, there needs to be one centralized oracle which makes channel allocation decisions for the entire network on the basis of prior channel environment conditions received from RNs. In contrast, in a distributed approach, the RNs need to make their own individual decisions in accordance with their local channel environment. The centralized approach requires complete information and coordination from/with RNs to provide a one-shot (static) optimal solution, which are more likely invalid by the time the central oracle decides them [6], [7]. Such an approach is impractical and not suitable for a time-sensitive and dynamically-changing network. Hence, our preference leans towards using the distributed decision approach to solve the problem of multi-band channel allocation to RNs providing reliable, timely, and valid decisions.

Our previous work considered distributed algorithms [6], [7] to solve this problem under the assumption of stable channel conditions and static radios. In this paper, we consider a dynamic relay network scenario with unstable channel environments and mobile radios. This complex, highly dynamic relay network environment requires a new distributed solution. We reformulate the original, computationally hard problem as a Markov decision process (MDP) [8]. Since solving the MDP to derive optimal multi-band channel allocation delay by individual RN is still an expensive process, we design its corresponding reinforcement learning problem, which is then solved using a customized Q-Learning algorithm with epsilon-greedy policy [9].

Contributions: With the objective of minimizing the packet latency in multi-band relay network, the contributions of our work, in this paper, are as follows.

1) We formulate an optimization problem for multi-band channel allocation under mobile and dynamic radio environment, and demonstrate its computationally hard nature and need for complete information. We then model the original problem as an MDP, and provide details on how the proposed model satisfies MDP properties.
2) We design this proposed MDP as a reinforcement learning-based problem where we provide details on designed states, actions and rewards. We propose a customized Q-learning algorithm to solve the reformulated problem.
3) We provide the convergence bounds of the proposed model through empirical results by fine-tuning the relevant learning parameters. We also compare the empirical performance of proposed method with comparable, conventional methods.

The remainder of the paper is structured as follows. In Section II, we describe the relevant research done in this field and its shortcomings. In Section III, our system

model is presented along with an optimization problem formulation. In Section IV, we provide preliminaries on MDP and reformulate the original optimization problem into an MDP. Section V presents our proposed Q-Learning-based reinforcement learning algorithm. The performance of our proposal is evaluated and compared with conventional methods in Section VI. Finally, the paper is concluded in Section VII.

II. RELATED WORK AND MOTIVATION

Improving the Quality of Service (QoS), particularly packet latency, is a trendy research topic in the area of wireless relay topologies. Various works have considered the incorporation of multiple frequency bands with multi-channel environment for the formation of a versatile relay network [10]–[16] with the objective of improving end-to-end packet delay.

A network topology comprising multiple hops was considered by the researchers in [10], and was constructed by centrally formulating an optimization problem. The work aimed to provide route and link scheduling along with channel assignment. Their research findings demonstrated that uncertain packet scheduling, arising from dynamically varying channel conditions, leads to system overload and disruption in the relay-based communication. However, scalability was a key issue in this work, mainly due to the assumption made by the central node that the wireless link conditions for a specific range do not vary during a specific time period.

On the other hand, researchers in [11] designed a multi-band relay topology to mitigate packet latency in wireless communications in which simultaneous reception and transmission of packets were performed at each RN. In contrast to tradition methods operating with the complete header information, this work depended on a truncated header. Thus, while receiving data packets, a RN relays (re-transmits) the packets to the subsequent node right after reading the truncated header instead of reading it completely. This strategy is called the truncated decode and forward (TDF) scheme. To obtain the information for assessing whether the relay transmission is required or not, the RN receives the header of a frame over one frequency band. If relay transmission is required, the RN commences the relay transmission on another frequency channel while the RN continues receiving the frame [11]. However, reading the partial header results in increased frame error probability since it can degrade the decoding performance. As a result, the likelihood of packet re-transmission also increases, which in turn adversely impacts the packet latency.

The research done in [12] provided empirical results using a WLAN system making use of multiple bands in tandem. By considering the Cognitive Radio Network (CRN) type white spaces in the spectra spanning the multiple frequency bands, this approach significantly improved the spectral efficiency. However, this work only considered algorithm design for the simultaneous use of multiple bands in a short-range communication between a pair of nodes, and did not extend this approach to a relay-based topology due to various challenges.

The relay topology was considered by researchers in [13], who aimed to avoid packet loss due to the gap in data rates used by SN and RN nodes. To combat this problem, an algorithm was designed by considering Signal-to-Noise Ratio (SINR) information when receiving or sending (relay) the packets. The acceptable interference had to be fine-tuned in the algorithm to optimize the spatial reuse by selecting channels that can maximize the reception/relay transmission rates while minimizing the packet forwarding delay. In this way, the SINR associated with every channel between SN-RN/RN-RN/RN-DN was considered to be known so that the best modulation and coding scheme (MCS) could be selected accordingly. However,the shortcoming of this work is its inability to handle the scenario where better SINR channels are not available.

Congestion-awareness was considered to mitigate the impact of interference between SN and RN nodes by researchers in [15]. The congestion-awareness algorithmic design was also inspired by the need to improve the residual energy performance of the battery-powered RNs so that their operational lifetime could be prolonged. The designed algorithm evaluated congestion levels on the various channels and selected the least congestion-prone channel. While this approach helped improve the aggregate throughput, its computational burden due to the brute-force iteration was not considered in the work. In a similar spirit to ranking the channels according to their performance level, researchers in [16] considered the available channel properties of all nodes. Then, a distributed channel allocation algorithm was conceptualized by permitting each node to select the most appropriate channel, based on its residual energy as well as the channel rank. In addition, spectrum-sensing and sleep duration optimization were jointly performed to satisfy the energy consumption constraints and increase throughput. Compared with the aforementioned approach, this technique demonstrated that distributed channel allocation is, indeed, possible without any central coordination or global information.

While such methods typically rely on pre-established rules, the highly dynamic network conditions make it difficult to manually design and incorporate any pre-defined rules in a hard-coded manner. To address this issue of mobile networks, machine learning (ML) techniques have been extensively considered in the literature which, has also motivated us to explore and propose ML-based algorithm in this paper. For instance, reinforcement learning has been previously employed in different types of cognitive radio networks (CRNs) for adaptive access of the dynamic spectrum. Felice *et al.* presented a survey as well as a proof-of-concept for the reinforcement learning-assisted spectrum management of CRNs [17]. The categories of the spectrum-sensing problem were further investigated in [18], revealing that the sequential multi-channel selection in CRNs leads to sensing-order and stopping rule problems. On the

other hand, periodic and channel-specific spectrum sensing was identified to be a cooperative sensing scheduling (CSS) problem leading to an under-utilization of the available spectrum. However, this work pointed out that a distributed channel allocation in a CRN setting is recommended; however, it is challenging due to the aforementioned problems. That said, none of the aforementioned work investigated the coexistence of multiple frequency bands for the channel allocation problem, let alone taking into account the relay-based topology and user/relay movement in emerging networks. This proves the novelty of our proposed reinforcement learning-based ML algorithm, which provides a distributed self learning channel allocation solution for highly dynamic multi-band relay network.

The work in [19] demonstrated how different frequency bands having various propagation properties, particularly in the millimeter (mmWave) bands, affects the hand-off mechanism due to the need to select the best base-station via the best possible frequency band. The UE-base-station hand-off policy is developed as a MDP to achieve the optimal hand-off decision by considering the remaining bandwidth of the serving base station, link conditions, and the user's maximum pay per connection budget.

The lessons learned from the literature review are summarized in this paragraph, depicting the motivation behind our proposed algorithms and presenting the novelty of this research work: The centralized decision algorithms are typically computationally hard and require non-deterministic polynomial time. The use of such approaches in literature, fail to consider the rapidly changing channel conditions. This motivates us to explore distributed self learning algorithm, known to be ML techniques, in order to propose smart and adaptive decision approach. On the other hand, the management of a limited buffer at RNs is also not considered in many existing research works, even though they can have a detrimental effect on the packet latency in the relay network. Furthermore, corner cases involving high and low SNR, impacting throughput and delay performance of RNs, are also not taken into account in the existing literature. Hence, a system model to address these various aspects is required, which we present in the following sections of this paper.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we describe a system model for a multi-band relay-based network. We then present an optimization problem that considers the assignment of variable-frequency band channels to the RNs. For ease of reference for the readers, the major notations and symbols used throughout this section and the remainder of the paper are listed in Table 1.

Before we begin, a few assumptions need to be stated. First, some mobile terminals have already been selected as RNs, and the incentive/policy used by the network operator allowing some mobile users to act as RNs is beyond the scope of our current work. Second, spectrum sensing [20], [21] is assumed to be performed by the radios before accessing channels. Third, the Signal-to-Noise Ratio (SNR) estimation [22] is assumed to be carried out by the nodes before taking decisions. Any RN selection approach [23], spectrum sensing algorithm, and SNR estimation technique can be used for the practical implementation of proposed algorithm. Note that the joint channel equalization and estimation for multiple frequency bands is still an open research issue. Since this is also beyond the scope of this research, we consider the channel estimation information to be available to the RNs, because this is assumed to have a constant and relatively negligible effect on the packet latency in our relay network.

It is to be noted that, subscripts *sr* and *rd* are used for any pair of SN-RN (where RN receives data from SN) and RN-DN (where RN needs to forward the same data packet to DN), respectively. Moreover, for any particular RN, its previous node is said to be the SN and the node ahead of RN is said to be the DN. Once any RN starts receiving a data packet F_{sr}, from its previous SN, at time t_{ir} and channel f_{sr}, with a data rate $\mathcal{D}_{sr}$, the RN, being able to communicate on multi-bands, can now commence forwarding the data packet F_{rd} to its DN, at t_{out} at data rate $\mathcal{D}_{rd}$, using some another channel f_{rd} from a different band. At this point,

TABLE 1. List of major technical notations and symbols used in the paper.

Notations	Description
F_{sr}	Data frame sent from SN to RN in bits.
F_{rd}	Data frame from RN to DN in bits.
t_{ir}	Time at which RN starts receiving F_{sr}
t_{out}	Time at which RN schedules the data frame for re-transmitting it to DN (F_{rd}).
T_h	Time span of the header of the packet.
T_{sr}	Time span of F_{sr}.
T_{rd}	Time span of F_{rd}.
T_{tot}	Total time span from the instant at which RN starts receiving data frame from SN (F_{sr}) till the time instant, DN completely receives the re-transmitted data frame (F_{rd}) from RN.
$\mathcal{D}_{sr}$	Rate at which SN transmits data frame to RN (F_{sr}).
$\mathcal{D}_{rd}$	Rate at which RN re-transmits data frame to DN (F_{rd}).
ch_{sr}^{ij}	i-th frequency channel of j-th frequency band used by SN to transmit F_{sr} to RN.
ch_{rd}^{mn}	m-th frequency channel of n-th frequency band used by RN to re-transmit F_{rd} to DN.
$\mathrm{SNR_{sr}}$	Estimated SNR of the channel used by SN to transmit F_{sr} to RN.
$\mathrm{SNR_{rd}}$	Estimated SNR of the channel used by RN to re-transmit F_{rd} to DN.
buffsize	Buffer size available to the node.
f_{rd}	Frequency channel used by the RN to forward data packet to its next destined node.
f_{sr}	Frequency channel used by the source node to transmit data packet to the RN.
α	Learning rate of Q-Learning algorithm.
γ	Future reward parameter for future state/action.
ϵ	Parameter for probability of choosing a random action. $(1-\epsilon)$, is probability of choosing greedy (best) action).

the RN needs to decide the feasible $\mathcal{D}_{rd}$, t_{out} and f_{rd} values suitable for the incoming data rate. The selection of $\mathcal{D}_{rd}$, t_{out} and f_{rd} depends on the resources available at the RN that include available power P_{avail}, size of the buffer (buffsize), and availability/occupancy of frequency channels.

In the case where RN finds that all channels are occupied, it forwards packet using DF scheme (that is on the same channel, $f_{rd} = f_{sr}$) by finding the best possible $\mathcal{D}_{rd}$. In this case:

$$T_{tot} = T_{sr} + T_{rd}, \tag{1a}$$
$$t_{out} = t_{ir} + T_{sr}, \tag{1b}$$
$$BO = F_{sr} - \text{buffsize}, \tag{1c}$$

where, $T_{sr} = F_{sr}/\mathcal{D}_{sr}$ and $T_{rd} = F_{rd}/\mathcal{D}_{rd}$ [6], defined as: the packet time span calculated as frame size divided by data rate as described in Table 1. T_{tot} denotes the total time needed for RN to complete its simultaneous reception and re-transmission of F_{sr} and F_{rd}, respectively. BO represents the observed buffer overflow in cases where buffsize is limited.

In the case where the RN finds the available channels with an SNR better than that of the receiving channel ($\text{SNR}_{\text{rd}} > \text{SNR}_{\text{sr}}$), the data packets can now be forwarded at a higher data rate ($\mathcal{D}_{rd} > \mathcal{D}_{sr}$). This results in $T_{rd} < T_{sr}$. In this case [6]:

$$T_{tot} = T_{sr}, \tag{2a}$$
$$t_{out} = t_{ir} + T_{sr} - T_{rd}, \tag{2b}$$
$$BO = [(T_{rd} - (T_{sr} - T_h))\mathcal{D}_{rd}] - \text{buffsize}, \tag{2c}$$

where T_h is the header time of F_{sr}. When the channel has lesser SNR_{rd} than SNR_{sr}, the packet can be forwarded at a lesser $\mathcal{D}_{rd}$ than $\mathcal{D}_{sr}$. This results in $T_{rd} > T_{sr}$. In this case [6]:

$$T_{tot} = T_{rd} + T_h, \tag{3a}$$
$$t_{out} = t_{ir} + T_h, \tag{3b}$$
$$BO = [(T_{sr} - T_{rd})\mathcal{D}_{sr}] - \text{buffsize}. \tag{3c}$$

Considering this system model, we can now formulate the multi-band channel assignment to the RNs as a minimization problem with the following pseudo equations, where Eq. (4a) is the objective function, under several constraints (Eqs. 4b, 4c, 4d). Note that constraint Eq. (4c) signifies the highly dynamic relay conditions since the SNR of the links between the RNs (or relay-destination nodes) is treated as a function of the link power, frequency bands dynamics, and the changing distance under the effects of mobility.

$$\min_{\mathcal{D}_{rd}, f_{rd}} \quad (t_{out} + T_{rd}), \tag{4a}$$
$$\text{s.t.} \quad BO(\mathcal{D}_{sr}, \text{buffsize}) \leq \text{BO}_{\text{min}}, \tag{4b}$$
$$\text{SNR}_{\text{rd}}(\text{P}_{\text{rd}}, \text{f}_{\text{rd}}, \text{dist}_{\text{rd}}) \geq \text{SNR}_{\text{min}}, \tag{4c}$$
$$P_{rd} \leq P_{avail}, \tag{4d}$$

where $BO(\mathcal{D}_{sr}, \text{buffsize})$ is the observed buffer overflow being a function of $\mathcal{D}_{sr}$ and buffsize (Eqs. 1c, 2c, 3c), BO_{min} is the minimum allowed value of overflown buffer, $\text{SNR}_{\text{rd}}(\text{P}_{\text{rd}}, \text{f}_{\text{rd}}, \text{dist}_{\text{rd}})$ is the SNR on certain channel being a function of P_{rd}, f_{rd} and $dist_{rd}$ Eq. (10), P_{rd} is the power required to transmit at $\mathcal{D}_{rd}$ and $dist_{rd}$ is the distance between the RN and its next DN and SNR_{min} is the QoS constraint.

Based on our earlier work in [7], we can treat this optimization problem as a computationally hard (NP) problem that cannot be solved in polynomial time for a relay network with a large number of RNs, frequency bands, and channels. Furthermore, complete information across the entire relay topology is required by a central oracle to solve this problem even for a relatively small search space. Due to these practicality issues, we explore how to remodel this optimization problem in a distributed manner in the following section.

IV. PROBLEM REFORMULATION AS MDP

In this section, we aim to reformulate the original optimization problem Eq. (4) as an MDP from the perspective of a certain RN. This is required for a distributed optimal decision regarding the multi-band channel allocation at the RN level, where we are given only the local information available to the RN. We first provide the preliminaries on MDP, and then present the problem reformulation as an MDP.

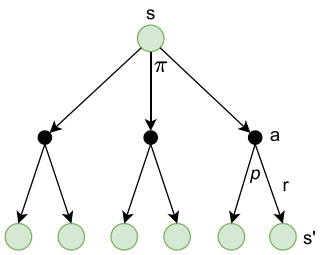
A. MDP PRELIMINARIES

Markov processes are characterized by having future states that depend only on the current state. For example, if at a certain time, an agent is in state S_t, a random action is being taken A_t, now the occurrence of the next state S_{t+1} depends on this current action, A_t, only and not on action taken at the previous instant, A_{t-1}. Above is the property of any process to be defined as a Markov process. Our proposed problem, i.e., the RN choosing making a decision about forwarding a data packet, is modeled as an MDP and is constructed in such a way that it can learn from the random actions taken in certain states.

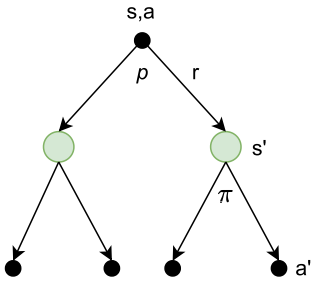
For any system, these actions and states have certain values, which are defined by Bellman's equations and can also be observed from backup diagrams as shown in Fig. 2. The following is the Bellman equation for the value of any state S following policy π, denoted as V_π:

$$V_\pi(S) = \sum_A \pi(A|S) \sum_{S'} p(S'|A)\big[R + \gamma V_\pi(S')\big], \tag{5}$$

where A is the action taken from state S, S' is the next state, γ indicates the future reward parameter to give importance to the value of the next state, $V_\pi(S')$. Through this equation, the values of a current and subsequent state(s) can be related as depicted in Fig. 2(a). This can be regarded as foreseeing the possible states in the future with respect to the current state, each of which is denoted by an open circle. The state-action pairs are presented as solid circles. As shown in Fig. 2(a), the top (root) node signifies state s, from which any of the three set of actions could be taken by adhering to a policy π. The environmental response could be a further sequence of



(a) Backup diagram for value of a state while following policy π.



(b) Backup diagram for value of action when following policy π.

FIGURE 2. Backup diagrams demonstrating the MDP for values of a state and an action, respectively, subject to a given policy.

succeeding steps that can be measured by a reward. The reward calculation is done by a function p. By averaging over all the possible states and weighing each state in terms of its probability, Eq. (5) ascertains that the beginning state value equals the sum of the expected next state and the expected reward.

Similarly, the following is the Bellman equation for the value of an action A taken according to a policy π from the state S, donated as $Q_\pi(S, A)$:

$$Q_\pi(S, A) = \sum_{S'} p(S'|A)\left[r + \gamma V_\pi(S')\right] \tag{6}$$

Imagine the solid circle at the top of Fig. 2(b) as an action taken from some state S. This action taken now has different probabilities p of ending up in either of the next two states shown in the figure. The Bellman equation for an action, averages over all the possibilities of its successor states and rewards, prioritizing each according to its likelihood. Thus, Eq. (6) represents the beginning state value as sum of the anticipated next state and its associated reward.

B. PROPOSED MDP MODEL

Based on the preliminaries, we can now formally transform our original problem Eq. (4) as a distributed, finite-state MDP as depicted in Fig. 3. Our designed states, actions, and rewards are presented in the figure. The MDP model comprises seven states, denoted by $\boldsymbol{S} = \{S_1, S_2, S_3, S_4, S_5, S_6, S_7\}$, where an RN is an agent taking actions and interacting with the environment under dynamically changing channel conditions and considering mobility.

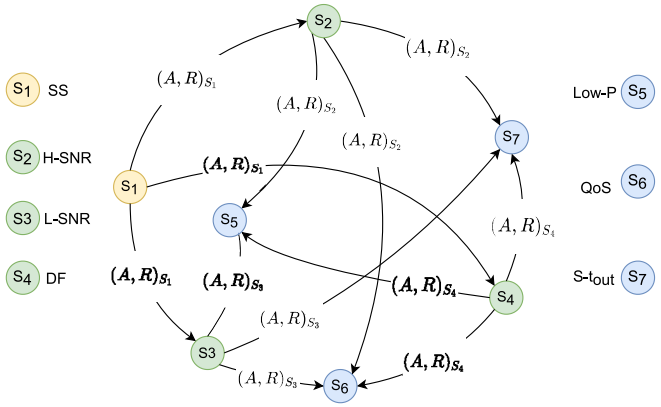


FIGURE 3. State diagram for proposed finite state MDP reinforcement learning algorithm.

The set of states, and their corresponding action sets, rewards sets and next state sets are discussed in detail as follows:

1) S_1 is the **initial state** of RN, where RN performs **spectrum sensing** along with **SNR estimation** on a certain frequency channel from any band. We refer this state as the **SS-state**.
Actions Set: The set of actions at this state comprises of set of frequency channels, to be sensed and estimated, given as: $\boldsymbol{A_{s_1}} = \{f_1, f_2, \ldots, f_{\mathcal{F}}\}$, where $\mathcal{F}$ presents the total number of channels that can be picked in order to be sensed. Hence, the number of possible actions available for RN in this state is equals to $\mathcal{F}$.
Rewards Set: The associated set of rewards for this state is given as $\boldsymbol{R_{s_1}} = \{R_1^{s_1}, R_2^{s_1}, R_3^{s_1}\}$, such that $R_1^{s_1} > R_2^{s_1} > R_3^{s_1}$. Here, rewards are being provided on the basis of sensing results which could be either available with good SNR, available with bad SNR, or not available at all. Once the RN pick say, f_1, and senses it as free with high SNR, the environment assigns $R_1^{s_1}$ against action f_1, leading the RN to the next state of S_2. In the case where RN finds this channel to be free, but with lower SNR, $R_2^{s_1}$ gets assigned to the action f_1 with the next state as S_3 and if the sensed channel is occupied, $R_3^{s_1}$ gets assigned to the action f_1, leading the agent to the next state of S_4. The set of next states for this state is given as $\boldsymbol{S'_1} = \{S_2, S_3, S_4\}$.
2) S_2 is an **intermediate state** of proposed MDP model, where RN has reached after finding out the initially (at initial state) picked frequency channel to be free with higher SNR. This state is referred to as the **H-SNR-state**.
Actions Set: In this state, RN takes an action of picking a data rate, achievable on this high SNR free channel. Hence, the set of actions for this state can be given as: $\boldsymbol{A_{s_2}} = \{\mathcal{D}_1^{s_2}, \mathcal{D}_2^{s_2}, \ldots, \mathcal{D}_{\mathcal{H}}^{s_2}\}$, where $\mathcal{H}$ denotes the total number of data rates achievable at higher SNR channels.
Reward Set: With reference to Fig. 4 and Eq. (2a), the data rates chosen in this state does not impact

the total time of simultaneous reception and transmission. Hence in this state, set of rewards are designed as: $\boldsymbol{R_{s_2}} = \{R_1^{s_2} \propto (\mathcal{D}_1^{s_2})^{-1}, R_2^{s_2} \propto (\mathcal{D}_2^{s_2})^{-1}, \ldots, R_{\mathcal{H}}^{s_2} \propto (\mathcal{D}_{\mathcal{H}}^{s_2})^{-1}\}$, where rewards are being assigned against actions (data rates) in an inversely proportional manner, giving higher reward to the lesser data rate, as it utilizes lesser power and lesser reward to higher data rate as it utilizes more power. In other words: in this state, the faster the data rate, the more the buffsize and the more the power is required to re-transmit data packet, without effecting total transmission time [6], hence in this state, we are assigning the least reward to the fastest data rate and the greatest reward to the slowest data rate. If the selected data rate requires power more than the P_{avail} of the RN, the environment assigns a zero or negative reward (also known as bad rewards) to the selected action and leads the agent to next state of S_5. On the other hand, if the selected data rate ends up violating the QoS, the environment drives the agent to state S_6 by assigning a bad reward to the action. However, if both constraints are met, the environment leads the agent to state S_7 by assigning rewards in accordance with set $\boldsymbol{R_{s_2}}$. So for this state, set of next states is given as: $\boldsymbol{S_2'} = \{S_5, S_6, S_7\}$.

3) S_3 is an **intermediate state** of the proposed MDP model, where RN has reached after finding out the initially (at previous state) picked frequency channel to be free with lesser SNR. This state is referred to as the **L-SNR-state**.
Action Set In this state, RN takes an action of picking a data rate, achievable on this low SNR free channel. The set of actions for this state can be given as: $\boldsymbol{A}_{s_3} = \{\mathcal{D}_1^{s_3}, \mathcal{D}_2^{s_3}, \ldots, \mathcal{D}_{\mathcal{L}}^{s_3}\}$, where $\mathcal{L}$ is the total number of data rates achievable at this state.
Rewards Set As can be seen from Fig. 4 and Eq. (3a), the data rates selected in this state has direct impact on the total time of simultaneous reception and transmission. Hence, in this state, set of rewards is designed as: $\boldsymbol{R_{s_3}} = \{R_1^{s_3} \propto \mathcal{D}_1^{s_3}, R_2^{s_3} \propto \mathcal{D}_2^{s_3}, \ldots, R_{\mathcal{L}}^{s_3} \propto \mathcal{D}_{\mathcal{L}}^{s_3}\}$, where rewards are being assigned to actions in a directly proportional manner, giving higher rewards to actions resulting in shorter total transmission time (higher data rates) and assigning lesser rewards to actions resulting in longer total transmission time (lower data rates). In other words: in this state, the faster the data rate, the lesser the buffsize and the lesser the total transmission time is required to forward the data packet [6], hence in this state, we are assigning the greatest reward to the fastest data rate and the least reward to the slowest data rate. In this state, after an action is taken, the environment follows the same criteria for choosing the next states as discussed in case of state S_2.

4) S_4 is an **intermediate state** of the proposed MDP model, where RN has reached after finding out the previously chosen frequency channel to be occupied/unavailable. This state is referred to as the **DF-state**.
Actions Set: Here, the actions set is a union set of all the data rates from states S_2 and S_3 given as: $\boldsymbol{A_{s_4}} = \boldsymbol{A_{s_2}} \cup \boldsymbol{A_{s_3}}$, depicting all data rates, achievable on the frequency channels once they get free at time t_{out} (Eq. (1b) and Fig. 5).
Rewards Set: With reference to Fig. 5 and Eq. (1a), here the selected data rate has direct impact on total transmission time. Hence, in this state, set of rewards is

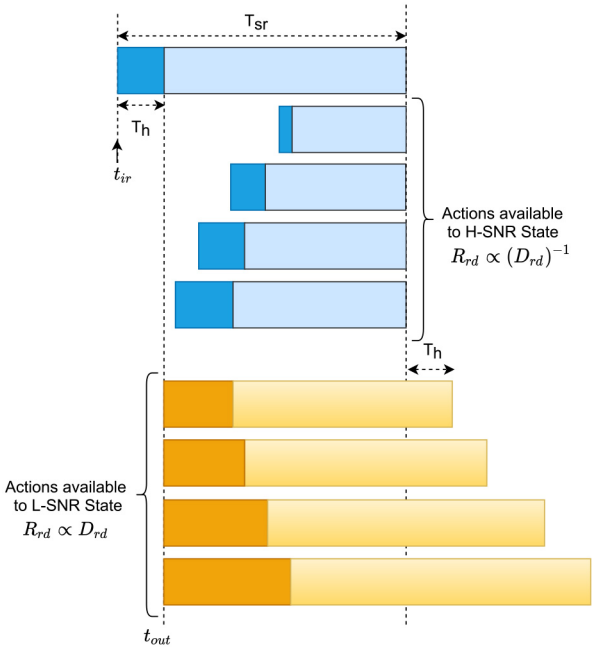


FIGURE 4. High-level view of action sets available to S_2 and S_3 and their corresponding rewards.

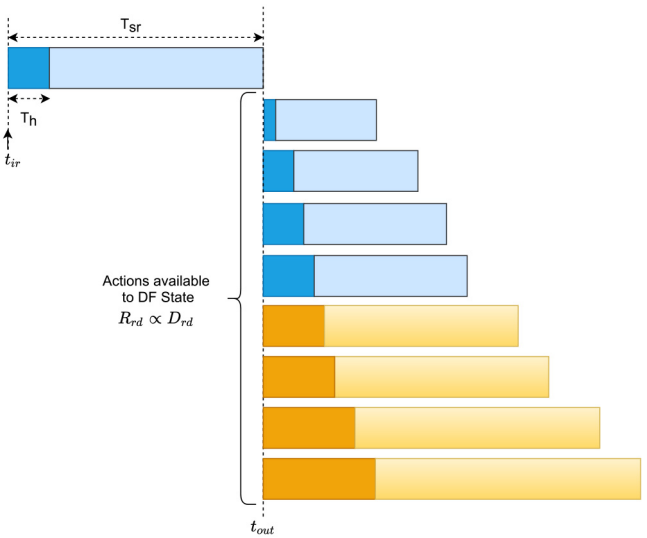


FIGURE 5. High-level view of action sets available to S_4 and its corresponding rewards.

given as: $\boldsymbol{R_{s_4}} = \{R_1^{s4} \propto \mathcal{D}_1^{s4}, R_2^{s4} \propto \mathcal{D}_2^{s4}, \ldots, R_{\mathcal{L}+\mathcal{H}}^{s4} \propto \mathcal{D}_{\mathcal{L}+\mathcal{H}}^{s4}\}$, where rewards are being assigned to actions in a directly proportional manner, giving higher rewards to actions resulting in shorter total transmission time (faster data rates) and assigning lesser rewards to actions resulting in longer total transmission time (slower data rate). In other words, in this state, the faster the data rate the lesser the total transmission time is required to forward the data packet [6], hence in this state, we are assigning the greatest reward to the fastest data rate and the least reward to the slowest data rate. In this state, after an action is taken, the environment follows the same criteria for choosing the next states, as discussed for state S_2.

5) S_5 is the **bad terminal state** of the proposed model, where RN reaches from any of the states, S_2, S_3, S_4 after taking an action (choosing data rate) requiring power more than P_{avail}. We refer to this state as **Low-P-state**. The actions causing this state, are given poor/bad rewards to assist the further learning process.
6) S_6 is the **bad terminal state** of the proposed model, where RN reaches from any of the states S_2, S_3, S_4 after taking an action (choosing data rate) resulting in SNR lower than the minimum QoS requirement, SNR_{min}. We call this state the **QoS-state**. The actions causing this state are given poor/bad rewards, to further help the learning process.
7) S_7 is the **good terminal state** of the proposed model, where RN reaches from any of the states S_2, S_3, S_4 after taking a feasible action (choosing a feasible data rate), in terms of P_{avail} and SNR_{min}. The actions causing this state are given rewards according to their corresponding reward sets $\boldsymbol{R}_{s_2}, \boldsymbol{R}_{s_3}, \boldsymbol{R}_{s_4}$. We call this state the **S-t_{out}-state**. At this state, the data packet gets scheduled at t_{out}, for re-transmission to the next node.

While the proposed MDP-based reformulated problem model is theoretically elegant, estimating the transition probabilities between the states and resolving an optimal channel assignment is not trivial, and remains computationally expensive for an individual RN. Therefore, a distributed learning technique is needed to solve the MDP for optimal multi-band allocation to RNs, which we design in the following section.

V. PROPOSED REINFORCEMENT LEARNING-BASED OPTIMAL MULTI-BAND ALLOCATION TO RNS

To solve the re-formulated MDP problem, we aim in this section to design a reinforcement learning algorithm to capture the ongoing process of interaction between an agent (i.e., a RN) and its environment. The agent being in a certain state, S, takes an action, A, according to some policy, and it consults with its environment as shown in Fig. 6. According to the action taken by the agent, the environment now assigns a particular reward, R, against that action and also decides the next state, S', to which the agent can move. During these

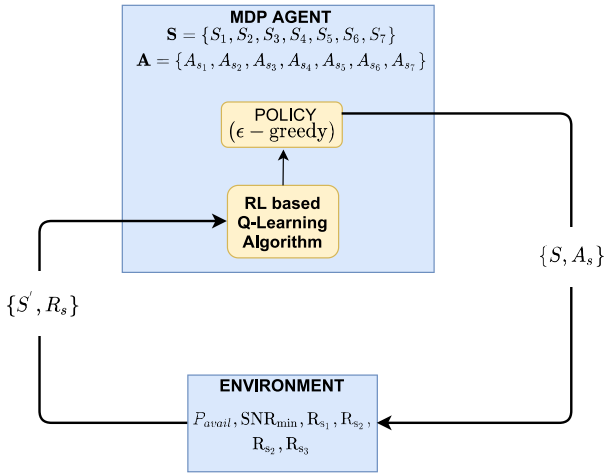


FIGURE 6. A high level illustration of the finite MDP reinforcement learning model to solve the reformulated problem.

continuous interactions, the RN acts as the agent learning the best sates and actions for itself. The actions are taken on the basis of a certain policy that could be either totally random, greedy or epsilon-greedy (ϵ-greedy). The transition from one state to the next has certain probabilities, known as transition probabilities (p) and which can be written in transition tables [9]. We represent the reformulated problem as a finite MDP reinforcement learning model, which consists of a stochastic 4-tuple of states, actions, rewards and transition probabilities (S, A, R, p).

Next, we detail the proposed distributed Q-Learning algorithm following an ϵ-greedy policy. In order to find the optimal value of the previously discussed Bellman equations (Eqs. 6) and (5) of any MDP, the Q-Learning algorithm can be used for finding the optimal decision solutions. In MDP, each and every possible action is first taken so the probabilities of their occurrence can be determined using Bellman equations, and then the final optimal value of any state and action can be found. On the other hand, Q-Learning updates the estimates, based on other (already) learned estimates, without waiting for a final outcome, a process also known as bootstrapping. Eq. (7) is known to converge to the optimal value of Eq. (6) for any MDP reinforcement learning problem.

$$Q(S, A) = Q(S, A) + \alpha\left[r + \gamma \max_{A} Q(S', A) - Q(S, A)\right] \quad (7)$$

where α represents the learning parameter, which impacts how fast the algorithm must learn. There are mainly three policies which can be followed when selecting an action, i.e., random, greedy, and ϵ-greedy [9]. In the random policy, the agent takes an action randomly at each state, thereby exploring the options. In the greedy policy, the agent always takes the best action (i.e., the one with the highest Q(S, A) value), thus exploiting the best known option. On the other hand, in the ϵ-greedy policy, the agent strikes a balance between exploring and exploiting the options by selecting the best

action using probability $(1-\epsilon)$ and taking a random action with probability ϵ. To ensure the best learning over time, we design our reinforcement learning solution by employing the ϵ-greedy policy to customize Q-learning as follows:

Back-end continuous Q-Learning algorithm: We utilize the Q-Learning algorithm as a continuous learning process for dynamic channel environment and mobile RNs. Referring to the steps provided in back-end Algorithm 1, there exists a Q-table for each state, containing Q-values against each action of that particular state. First, RN initializes all of its Q-values for every state. Then, the learning process starts where, RN has to update its Q-values for every state Eq. (7), in an episodic manner. For each episode RN, being a mobile node with varying channel conditions, first updates its $dist_{rd}$, its spectrum sensing results and also calculates its updated SNR estimation results. Then, RN takes an action (using afore-discussed epsilon-greedy policy) from the initial state, S_1, and reaches till its terminal state, following the designed MDP as shown in Fig. 3. For each episode, the path/sequence of the states from S_1 till any terminal state, depends on the description provided in Section IV-B. In each episode, RN runs its MDP (Fig. 3) and updates its corresponding Q-values, for every visited state, on the basis of its changed/updated channel environment and location. Utilizing famous Shannon [24] and free space path loss equations [25], the effect of mobility and changing channel conditions on the specified system can be observed from following equations [21]:

$$T_{rd} = \frac{\mathrm{F_{rd}}}{\mathcal{D}_{rd}}, \tag{8}$$

$$\mathcal{D}_{rd} = \mathrm{BW}\log_2(1+\mathrm{SNR_{rd}}), \tag{9}$$

$$\mathrm{SNR_{rd}} = \frac{\mathrm{P_{rd}h_{rd}}}{\mathrm{noise}}, \tag{10}$$

$$h_{rd} = \frac{c}{4\pi f_{rd} dist_{rd}}, \tag{11}$$

where c is the speed of light, BW is the bandwidth, h_{rd} is the path loss function, $dist_{rd}$ denotes the parameter getting affected with the mobility of the RN. As the RN changes its location, the distance between itself and its DN needs to be updated, which is being handled before every episode. The parameters $dist_{rd}$ and $noise$ directly affect the selection of suitable $\mathcal{D}_{rd}$ and f_{rd}. This is a continuous learning process which keeps running in the background. Once the RN learns its environment from running the episodes multiple times, Eq. (7), converges to the optimal decisions of $\mathcal{D}_{rd}$ and f_{rd} for dynamic mobile environment.

Q-Value exploiting algorithm: After getting trained by back-end Algorithm 1, RN's each MDP state, now, knows its converged optimal action, i.e., the actions having the highest Q-value. In real-time, at the reception of packet from prior node, these optimal actions get exploited by RN in order to finally decide the optimal channel for packet re-transmission. Through the steps of Algorithm 2, it can be seen that once the packet is received, the RN fetches the optimal action from its initial state's Q-table, unlike the back-end algorithm where the system first has to perform an entire learning procedure so as to reach to its decisions. The RN, continues exploiting optimal actions from every state till it reaches its terminal state, where the optimal channel can now be allocated for packet forwarding.

Algorithm 1: Backend Q-Learning Algorithm

Input: α, γ, ϵ, episodes
Output: $Q^*(S_1, A)$, $Q^*(S_2, A)$, $Q^*(S_3, A)$, $Q^*(S_4, A)$
for $S = S_1, S_2, S_3, S_4$ **do**
  for *all actions* **do**
    Initialize $Q(S, A)$ *with any value*
for $S = S_5, S_6, S_7$ **do**
  for *all actions* **do**
    Initialize $Q(S, A)$ *of terminal states with zero*
for *each episode* **do**
  Initialize $S = S_1$
  Use current location for $dist_{rd}$
  Use current channel situation for SS and SNR estimation
  for *each step of episode* **do**
    Choose action from available action set with ϵ*-greedy policy*
    Send to environment
    Observe corresponding reward (R) and next state (S')
    $UpdateQ(S, A) = Q(S, A) + \alpha[r + \gamma \max_A Q(S', A) - Q(S, A)]$
    Return S'
  Return once S==terminal state

Algorithm 2: Current Time Optimum Q-Value (Q^*) Exploiting Procedure for Optimal Multi-Band Channel Allocation to RN

Input: F_{sr}, t_{ir}, $\mathcal{D}_{sr}$, $\mathrm{SNR_{sr}}$ buffsize, f_{sr}, $dist_{rd}$, $\mathrm{SNR_{min}}$, P_{avail}
Output: t_{out},$\mathcal{D}_{rd}$, f_{rd}
Save t_{ir}, the instant of receiving F_{sr}
for $S = S_1, S_2, S_3, S_4$ **do**
  Get current value of $maxQ(S, A)$ from back-end learning machine (Algorithm 1)
Select f_{rd} from action space of S_1 having $maxQ(S_1, A) == Q(S_1, f_{rd})$
Get S' corresponding to action f_{rd}
Select $\mathcal{D}_{rd}$ from action space of S' having $maxQ(S', A) == Q(S', \mathcal{D}_{rd})$
if $S' == S_2$ **then**
  $t_{out} = t_{ir} + T_{sr} - T_{rd}$
elseif $S' == S_3$ **then**
  $t_{out} = t_{ir} + T_h$
elseif $S' == S_4$ **then**
  $t_{out} = t_{ir} + T_{sr}$
return t_{out}

TABLE 2. Considered simulation parameters.

Parameters	Values
F_{sr},F_{rd}	3000 (bits)
BW	20 MHz
$\mathrm{SNR}_{\min}$	-5 dB
Standard data-rate set	6, 9, 12, 18, 24, 36, 48, 54 (Mbps)
Number of RNs	4 nodes
buffsize	[400, 600, 200, 800] (bits)
P_{avail}	[1, 3, 5, 4] (Watts)
Number of episodes	2000
ϵ, γ, α	0.1, 0.5, 0.1

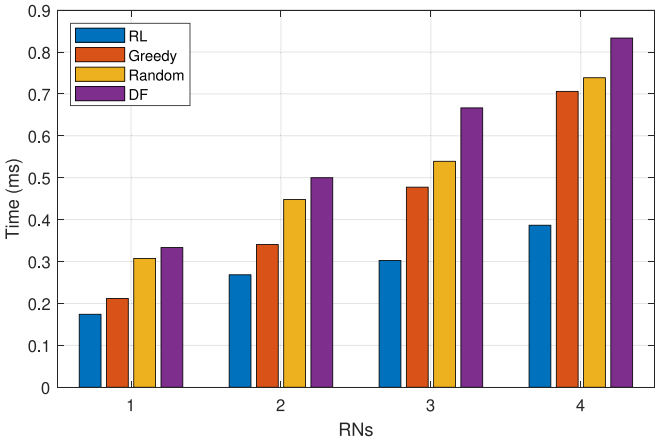


FIGURE 7. Comparison of the total transmission time for the considered methods with a varying number of RNs.

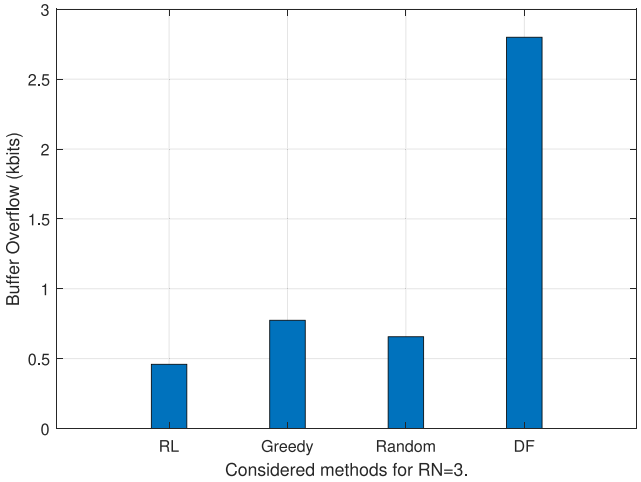


FIGURE 8. Buffer overflow comparison of RN = 3.

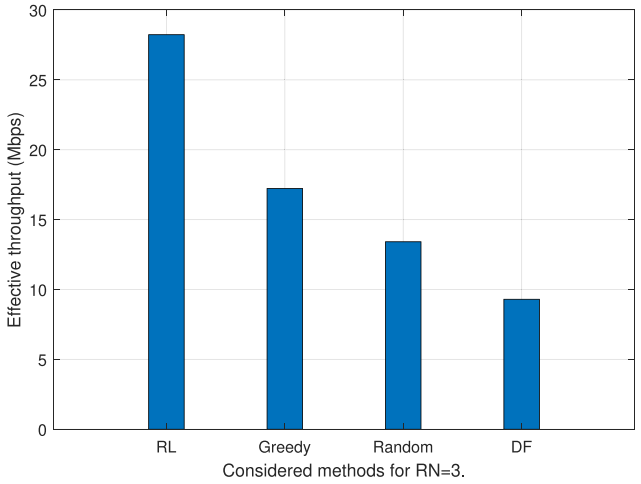


FIGURE 9. Effective throughput comparison of RN = 3.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposed reinforcement learning approach, based on computer-based simulations. First, we compare our proposal with conventional schemes (i.e., DF, random, and greedy channel selection [6] algorithms). Then, we also provide the effect of varying Q-learning parameters α and γ on the learning process. MATLAB scripts [26] are used to construct simulations using the parameters given in Table 2. The starting data rate from the main SN is set as $\mathcal{D}_{sr} = 18$ Mbps with $\mathrm{SNR}_{\mathrm{sr}} = 8$dB for plotting the results.

First, in Fig. 7, our proposed Q-Learning-based reinforcement learning method is compared with the conventional DF scheme, the random channel assignment, and the greedy-heuristic algorithms [20] for the worst case scenarios where only low SNR channels are available with respect to the incoming packet's channel. It can be noticed from the results that using the conventional DF scheme (without multi-band), the total transmission time delay is the longest because the RNs are receiving and transmitting the data packets one at a time. In the random channel allocation, the RNs arbitrarily select channels, resulting in an unfeasible channel selection, which results in the second longest time to relay the data packets to DN. Furthermore, it can be observed that the greedy algorithm performs somewhat better than those of DF and random channel selection. However, in the worst case scenario, the proposed reinforcement learning algorithm provides more feasible, converged results.

Next, in Fig. 8, we provide the buffer overflow results of our proposal in contrast with the conventional DF scheme, the random channel assignment, and the greedy-heuristic algorithm. It can be seen from the results that even in the worst-case scenario, the proposed learning model still converges so as to induce a minimum number of buffer overflow bits as compared with all other approaches. Also note that the buffer overflow of DF scheme is the highest, which corroborates the need for a multi-band channel selection method at the distributed (RN) level.

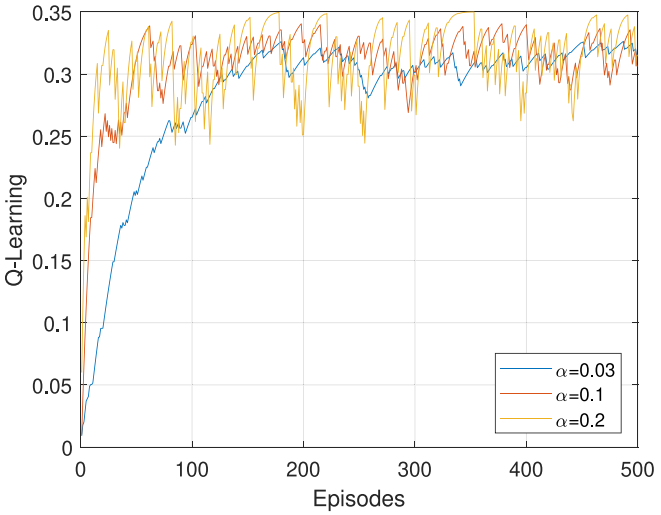


FIGURE 10. Effect of varying α with constant $\gamma = 0.1$.

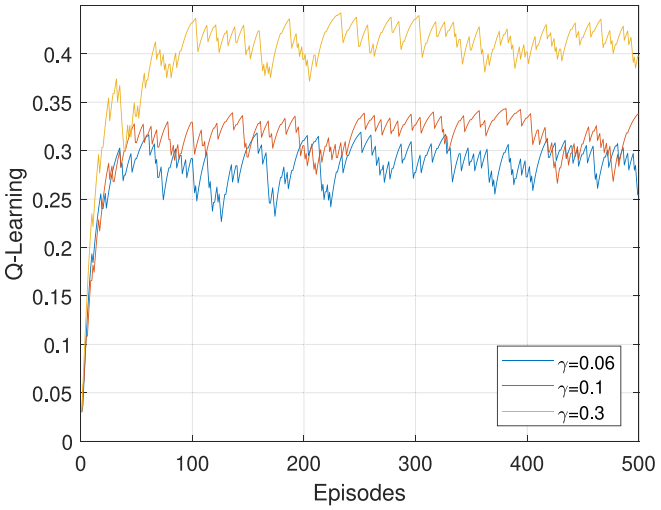


FIGURE 11. Effect of varying γ with constant $\alpha = 0.1$.

Next, we report the effective throughput performance of the compared methods. The effective throughput is defined as the number of successful packets received during total time of relaying. It can be seen from Fig. 9 that the effective throughput of the proposed reinforcement learning approach is the highest due to its ability to maximize the time efficiency while minimizing buffer overflow. On the other hand, the DF method suffers from the least effective throughput by exhibiting an exponential rise in the packet latency.

Fig. 10 demonstrates the effect of varying the learning rate, α, of our proposed reinforcement learning algorithm while keeping the future reward parameter, γ, constant. If the learning rate is too fast, the algorithm converges quickly, but it does not maintain a stable value for long. On the other hand, as the learning rate is made smaller, the algorithm takes longer to reach to the convergence, but gets stable faster.

Fig. 11 demonstrates the effect of varying the future reward parameter, γ, of the proposed reinforcement learning algorithm while setting the learning rate, α, constant. When the future rewards are given lower values, the algorithm struggles to reach the optimum value. On the other hand, the greater the value given to the future rewards, the more easily the algorithm reaches its optimum value.

TABLE 3. Q values of each action available to initial state S_1 for every 100^{th} episode.

	Episodes						
	1	**100**	**200**	**300**	**400**	**500**	**600**
A_{s_1}	$Q(S_1, A), S'$						
4.75 GHz	0.03,2	0.18,2	0.14,4	0.14,2	0.15,2	0.13,4	0.21,4
2.4 GHz	0,0	0,0	0,0	0,0	0.01,4	0.06,3	0.06,3
4.4 GHz	0,0	0.04,3	0.04,3	0.079,3	0.11,3	0.14,3	0.19,3
4.2 GHz	0,0	0.04,3	0.08,3	0.079,3	0.08,3	0.09,3	0.15,3
800 MHz	0,0	0.03,4	0.04,4	0.05,4	0.06,4	0.06,4	0.06,4
4.5 GHz	0,0	0.06,2	0.09,2	0.11,2	0.13,2	0.36,2	0.40,2

TABLE 4. Q values of each action available to intermediate sates S_2, S_3 and S_4 for every 100^{th} episode.

	Episodes						
	1	**100**	**200**	**300**	**400**	**500**	**600**
A_{s_2} **(Mbps)**	$Q(S_2, A), S'$						
24	-0.03,5	-0.23,5	-0.28,5	-0.29,5	-0.29,5	-0.29,5	-0.29,5
36	0,0	-0.23,5	-0.28,5	-0.29,5	-0.29,5	-0.29,5	-0.29,5
48	0,0	-0.23,5	-0.28,5	-0.29,5	-0.29,5	-0.29,5	-0.29,5
54	0,0	-0.24,5	-0.24,7	-0.29,5	-0.29,5	0.19,7	0.19,7
A_{s_3} **(Mbps)**	$Q(S_3, A), S'$						
6	0,0	-0.03,6	-0.04,6	-0.04,6	-0.04,6	-0.05,5	-0.06,6
9	0,0	0,0	0,0	0.02,7	0.04,7	0.05,7	0.09,7
12	0,0	0.03,7	0.06,7	0.06,7	0.06,7	0.08,7	0.08,7
18	0,0	0.04,7	0.07,7	0.11,7	0.16,7	0.16,7	0.21,7
A_{s_4} **(Mbps)**	$Q(S_4, A), S'$						
6	0,0	-0.02,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,6
9	0,0	-0.03,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
12	0,0	-0.02,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
18	0,0	-0.02,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
24	0,0	-0.01,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
36	0,0	-0.01,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
48	0,0	-0.01,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5
54	0,0	-0.02,5	-0.03,5	-0.05,5	-0.06,5	-0.07,5	-0.07,5

Finally, using Tables 3 and 4, a simulation example is presented. This example is for an individual RN with buffsize = 400, P_{avail} = 5W and the same common simulation parameters as in Table 2. Table 3 shows the Q values of each action available to the initial state S_1 for every 100^{th} episode. Table 4 shows the Q values of each action available to intermediate sates S_2, S_3 and S_4 for every 100^{th} episode.

It can be seen from the tables that as number of episodes increase, the agent (RN) learns and converges to its optimal action for each state. As shown in Table 3, for S_1, till 400^{th} episode, the RN is mostly inclined towards seeing, A_{s_1} = 4.75GHz channel and $S' = S_2$ as its best solution for S_1. But right after the 400^{th} episode, the RN converges to its actual optimum solution for S_1, which is A_{s_1} = 4.5GHz channel and $S' = S_2$.

Since the initial state S_1 has converged to its next state to be S_2, S_2 is now considered to be the optimal intermediate state. Which means now S_2 also needs to converge to its optimal action. As shown in Table 4, for S_2, by the 500^{th} episode, the RN has converged to $A_{s_2} = 54$Mbps data rate and $S' = S_7$, which is a good terminal state of the system.

VII. CONCLUSION

Relay-based networks have been increasingly adopted in a wide range of scenarios, from device-to-device networks to drone-cells for improved capacity and coverage. In 5G+ and 6G integrated networks, the role of relay-based networks will be different from their predecessors for various reasons, one of which is the introduction and incorporation of various frequency bands and their simultaneous use by source, destination, and RNs.

While high frequency bands provide much higher capacity, they are constrained with more stringent path loss and blocking. Even more, the channel conditions may drastically change in ultra-high frequency spectra in the upper GHz and THz level, such as visual light communication (VLC). Optimal channel allocation to these RNs to combat the real-time traffic load variation and other network dynamics including user mobility is shown to be a computationally hard problem, and attempting to solve which even for a low number of RNs requires a centralized oracle-like platform (e.g., a software defined network controller, or a central cloud server) to compute optimal channel allocation decisions. This is intractable for wireless RNs, which cannot wait to receive a central decision while network dynamics continue to change. This warrants a smart and distributed solution native to the RNs.

In this paper, we addressed this complex problem as an MDP, the optimal solution of which is also shown to be expensive. This motivated us to customize a reinforcement-learning method to solve the problem with near-optimal performance in real-time at the RNs. Convergence shows the fast-learning curve for the proposed approach. Comparative results also demonstrate that proposed reinforcement learning-based approach achieves comparable performance to that of the centralized benchmark and also outperforms several existing techniques in terms of packet transmission time, buffer overflow, and effective throughput.

REFERENCES

[1] X. Bu, C. Liu, Q. Yu, L. Yin, and F. Tian, "Optimization on cooperative communications based on network coding in multi-hop wireless networks," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2020, pp. 384–387.

[2] I. Chatzigeorgiou, "The impact of 5G channel models on the performance of intelligent reflecting surfaces and decode-and-forward relaying," in *Proc. IEEE 31st Annu. Int. Symp. Pers. Indoor Mobile Radio Commun.*, 2020, pp. 1–4.

[3] A. Ben-Bassat *et al.*, "10.5 a fully integrated 27dBm dual-band all-digital polar transmitter supporting 160MHz for WiFi 6 applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 180–182.

[4] S. Sakib, T. Tazrin, M. M. Fouda, Z. M. Fadlullah, and N. Nasser, "A deep learning method for predictive channel assignment in beyond 5G networks," *IEEE Netw.*, vol. 35, no. 1, pp. 266–272, Jan./Feb. 2021.

[5] S. Sakib, T. Tazrin, M. M. Fouda, Z. M. Fadlullah, and N. Nasser, "An efficient and lightweight predictive channel assignment scheme for multiband B5G-enabled massive IoT: A deep learning approach," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5285–5297, Apr. 2021.

[6] B. Mughal, Z. M. Fadlullah, and S. Ikki, "Centralized versus heuristic-based distributed channel allocation to minimize packet transmission delay for multiband relay networks," *IEEE Netw. Lett.*, vol. 2, no. 4, pp. 180–184, Dec. 2020.

[7] B. Mughal, Z. Fadlullah, M. M. Fouda, and S. Ikki, "Allocation schemes for relay communications: A multi-band multi-channel approach using game theory," *IEEE Sens. Lett.*, vol. 6, no. 1, pp. 1–4, Jan. 2022.

[8] Z. M. Fadlullah, C. Wei, Z. Shi, and N. Kato, "GT-QoSec: A game-theoretic joint optimization of QoS and security for differentiated services in next generation heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1037–1050, Feb. 2017.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[10] M. Li, S. Salinas, P. Li, X. Huang, Y. Fang, and S. Glisic, "Optimal scheduling for multi-radio multi-channel multi-hop cognitive cellular networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 1, pp. 139–154, Jan. 2015.

[11] N. Egashira, K. Yano, S. Tsukamoto, J. Webber, and T. Kumagai, "Low latency relay processing scheme for WLAN systems employing multiband simultaneous transmission," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[12] N. Egashira, K. Yano, M. Sutoh, A. Sugitani, Y. Amezawa, and T. Kumagai, "Experimental evaluation of adaptive simultaneous transmission timing control considering idle/busy probability for multi-band wireless LAN," in *Proc. Asia-Pacific Microw. Conf. (APMC)*, 2018, pp. 881–883.

[13] A. Hanyu *et al.*, "Adaptive frequency band and channel selection for simultaneous receiving and sending in multiband communication," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 460–463, Apr. 2019.

[14] Z. M. Fadlullah *et al.*, "Multi-hop wireless transmission in multi-band WLAN systems: Proposal and future perspective," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 108–113, Feb. 2019.

[15] L. Zhang, S. Gao, K. Wei, W. Zhang, and F. Yongxin, "Multi-channel allocation algorithm based on congestion avoidance in wearable wireless sensor network," in *Proc. 27th Wireless Opt. Commun. Conf. (WOCC)*, 2018.

[16] K. Ding, H. Zhao, X. Hu, and J. Wei, "Distributed channel allocation and time slot optimization for green Internet of Things," *Sensors*, vol. 17, no. 11, p. 2479, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/11/2479

[17] M. D. Felice, L. Bedogni, and L. Bononi, *Reinforcement Learning-Based Spectrum Management for Cognitive Radio Networks: A Literature Review and Case Study*. Singapore: Springer, 2018, pp. 1–38. [Online]. Available: https://doi.org/10.1007/978-981-10-1389-8_58-1

[18] X. Huang, T. Han, and N. Ansari, "On green energy powered cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 827–842, 2nd Quart., 2014.

[19] M. S. Parwez and H. Olufowobi, "Cost-constrained handoff in next generation heterogeneous wireless networks," in *Proc. 9th IEEE Annu. Ubiquitous Comput. Electron. Mobile Commun. Conf. (UEMCON)*, 2018, pp. 911–916.

[20] B. Mughal, S. Hussain, and A. Ghafoor, "Cooperative sub-carrier sensing using antenna diversity based weighted virtual sub clustering," *China Commun.*, vol. 13, no. 10, pp. 44–57, Oct. 2016.

[21] B. Mughal, S. Hussain, and A. Ghafoor, "Cluster-based cooperative sub-carrier sensing using antenna diversity-based weighted data fusion," *Arab. J. Sci. Eng.*, vol. 41, no. 9, pp. 3425–3439, 2016.

[22] W. Wang, Y. Shen, and Y. Wang, "Low-complexity non-data-aided SNR estimation for multilevel constellations," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 113–116, Jan. 2020.

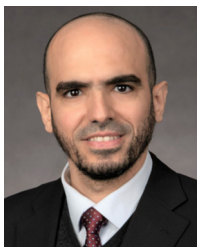
[23] J. Haghighat and W. Hamouda, "Decode-compress-and-forward with selective-cooperation for relay networks," *IEEE Commun. Lett.*, vol. 16, no. 3, pp. 378–381, Mar. 2012.

[24] J. Ma, "Modified Shannon's capacity for wireless communication [speaker's corner]," *IEEE Microw. Mag.*, vol. 22, no. 9, pp. 97–100, Sep. 2021.

[25] X. Fang, M. Ramzan, Q. Wang, N. Neumann, X. Du, and D. Plettemeier, "Path loss models for wireless cardiac RF communication," *IEEE Antennas Wireless Propag. Lett.*, vol. 20, no. 6, pp. 893–897, Jan. 2021.
[26] "MATLAB." [Online]. Available: https://www.mathworks.com (Accessed: Mar. 2022).

BUSHRA MUGHAL (Graduate Student Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Lakehead University, Thunder Bay, ON, Canada, in 2022. She has extensive experience in the development and deployment of software-defined radio-based wireless communication systems, as well as advanced knowledge of MATLAB/Simulink, LabView, GNU Radio, Python, USRPs, and NI VSTs. She has been engaged in comprehensive research in the areas of MIMO, OFDM, frequency hopping synchronization, spectrum sensing, cognitive radios, game theory, relay network, multiband communication, and machine learning. In addition to her numerous publications, she is currently interested in expanding her research area toward the optimization of wireless communication networks consisting of both ground and flying smart devices (UAVs/LAPs/NFP).

ZUBAIR MD. FADLULLAH (Senior Member, IEEE) is currently an Associate Professor with the Computer Science Department, Lakehead University, and a Research Chair of the Thunder Bay Regional Health Research Institute, Thunder Bay, ON, Canada. He was an Associate Professor with the Graduate School of Information Sciences, Tohoku University, Japan, from 2017 to 2019. His main research interests are in the areas of emerging communication systems, such as 5G New Radio and beyond, deep learning applications on solving computer science and communication system problems, UAV-based systems, smart health technology, cyber security, game theory, smart grid, and emerging communication systems. He received several best paper awards at conferences, including IEEE/ACM IWCMC, IEEE GLOBECOM, and IEEE IC-NIDC. He is currently an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IEEE Network Magazine*, IEEE ACCESS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and *Ad Hoc & Sensor Wireless Networks*. He is a Senior Member of the IEEE Communications Society.



MOSTAFA M. FOUDA (Senior Member, IEEE) received the Ph.D. degree in information sciences from Tohoku University, Japan, in 2011. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, USA. He is also an Associate Professor with Benha University, Egypt. He has served as an Assistant Professor with Tohoku University. He was a Postdoctoral Research Associate with Tennessee Technological University, Cookeville, TN, USA. He has published more than 90 papers in prestigious peer-reviewed journals and conferences. He has been engaged in research on cybersecurity, communication networks, wireless mobile communications, smart healthcare, smart grids, AI, blockchain, and IoT. He has served as a Guest Editor for several special issues of several top-ranked journals, such as IEEE WIRELESS COMMUNICATIONS and *IEEE Internet of Things Magazine*. He is an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Associate Editor of IEEE ACCESS.

SALAMA IKKI (Senior Member, IEEE) is an Associate Professor and the Research Chair of Wireless Communications with Lakehead University, Thunder Bay, ON, Canada. His research group has made substantial contributions to 4G and 5G wireless technologies. He is the author of more than 100 journals and conference papers and has more than 5500 citations and an H-index of 35. His group's current focuses on massive MIMO, cell-free massive MIMO, visible light communications, and wireless sensor networks. He received several awards for his research, teaching, and services.