Reinforcement Learning Based Vertical Handoff Decision Algorithm for Next Generation Wireless Network

Hemavathi and S. Akhila
B.M.S.C.E, Bengaluru and 560019, India
Email: hemavathi.ece@bmsce.ac.in; akhilas.ece@bmsce.ac.in

Abstract —Next-generation wireless network systems and technologies provide a new paradigm for achieving the fastest access to any network. But one of the significant design concerns is the support of handoff, irrespective of the services. The key objective of this work is to enable a node to make appropriate decisions for performing handoff through Reinforcement learning. The work concentrates on the handoff decision phase for choosing the best network with a minimum delay during the handoff process. The reduction in decision delay has been achieved by minimizing the number of handoffs. The environment is modeled as a Markov decision process with the aim of increasing the total anticipated reward per link. The network resources that are used by the link is taken by a reward function and network switching cost that is utilized to model the signaling and processing load incurred on the network during handoff. It has been shown that the total number of unnecessary handoffs can be decreased enhancing the performance of heterogeneous networks. Also, an assessment of the proposed scheme with the existing Vertical handoff decision algorithm like the Simple Additive Weighting method (SAW) has been made and the results show an improved performance over SAW.

Index Terms—Reinforcement learning, expected reward, vertical handoff, access point, value iteration algorithm, MDP.

I. Introduction

The Mobile Terminal (MT) of the 4th generation networks with various wireless interfaces is able to connect to various kinds of access networks. This has been possible due to the rapid improvement in technologies. In order to provide the always best connected (ABC) service continuity and consistent mobility to the user, the Mobile terminal needs to change its point of connection several times during the connectivity phase. This is called as the Handoff process [1]. Basically, there are two kinds of handoffs, namely horizontal and vertical (Fig. 1). The horizontal handoff is said to occur when the mobile terminals switch between the points of connectivity that use the same access technology. On the other hand, vertical handoff happens when the mobile terminals switch between the point of connectivity that use the different access technologies and sometimes, these connectivity points with different access technologies might sometimes be available in the same coverage area.

Manuscript received July 14, 2021; revised November 20, 2021.
Corresponding author email:hemavathi.ece@bmsce.ac.in
doi:10.12720/jcm.16.12.566-575

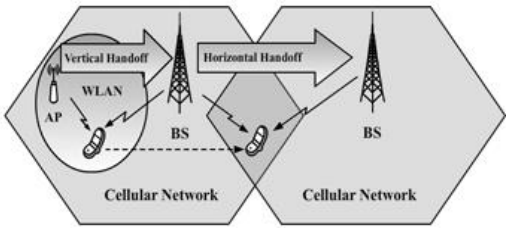

Fig. 1. Vertical handoff and horizontal handoff [2]

Vertical handoff takes place in three phases- system discovery, decision phase and execution phase [2]. A heterogeneous network basically means that different networks have different access technologies in terms of packet loss, bandwidth, latency, waiting time etc. Thus, it becomes important to make sure that the Mobile Terminals can work in the ABC mode without experiencing any delay that may be caused during the decision phase. In the Vertical handoff process, certain services and applications may be interrupted mainly because supporting handoff's across different networks with different access technologies with minimum delay is a challenge. This may be because each network might have its own handoff procedure and operational characteristics which results in unreasonable delay and packet loss. This could result in call blocking and call dropping issues impacting the connection level QoS directly. Thus, the problem of experiencing the handoff delay unnecessarily by the mobile terminal caused during the decision phase is intended to be solved by expressing it as a Markov decision process and then solving it using a Reinforcement learning based algorithm.

The rest of the paper is organized as follows: In Section II, work related to the most significant literatures focusing on enhancing the performance of heterogeneous network when vertical handoff takes place is presented. Section III presents the formulation of Markov Decision Process (MDP) model. Section IV presents the algorithm developed for reducing the delay experienced during handoff and also how the best network is selected using Reinforcement learning concept and in Section V result are discussed. Lastly, inferences are presented in Section VI.

II. Related Works

This section discusses about the existing techniques that has been introduced in the past for strengthening the vertical handoff mechanism. This section will update the findings from the prior review work

Rong Chai *et al*., [3], have described various vertical handoff decision algorithms centered on Received Signal Strength (RSS); cost functions that include network coverage area, bandwidth, cost of service, reliability factor, security, User Equipment (UE) mobility model, battery power etc. Also some of the tasks and glitches of the cost function and numerous characteristics decision making based handoff algorithms are discussed. It is observed that the handoff algorithm based on RSS alone is no longer appropriate for heterogeneous wireless network with various kinds of user services and priorities.

In [4] Y. Chen *et al*., proposed the handoff decision algorithm based on an MDP model used on the multimode terminal. The network's delay and available bandwidth are considered as key factors for developing the model. The multimode terminal can connect to various networks related to Multi-Domain (various operators) Heterogeneous Wireless Access Networks having different management strategies. A tradeoff amongst handoff cost and received QoS is made by using a reward function. An optimal policy for handoff is found by using a Markov decision method to define the problem with an aid to maximize the total reward over the period of transmission.

In [5], an optimized vertical handoff algorithm that included a combination of MDP and fuzzy logic method was developed. It enabled the mobile terminal to incorporate load balancing of a greater level that effectively minimizes the packet loss rate, average blocking rate and ping-pong effect. However, this is basically applied to vehicular heterogeneous network.

Enhancing the QoS and to reducing the number of handoffs through a dynamic network selection strategy has been addressed in [6]. The heterogeneous system model consists of LTE and UMTS which act as cellular networks and WiMAX, Wi-Fi act as the broadband access networks. A procedure based on Reinforcement Learning (RL) is developed that chooses the superlative network centered not only on the present network load but also on the possible future network states. It is found that the proposed network selection algorithm reduces the amount of handoffs obtained through the MDP algorithm, while maintaining high QoS.

In [7], A Vertical Handoff Decision algorithm based on User Mobility Pattern (UMP) in cellular-Wireless LAN is modeled as a MDP. The MDP states are defined by history records of Received signal strength (RSS) of WLANs. WLANs RSS records on the mobile terminal are used to derive the state transition probabilities which indicate user mobility patterns of specific WLAN Access Points. The recommended algorithm can decrease call dropping possibility intensely with little rise in the amount of handoffs and accomplish greater total expected reward. The proposed Vertical Handoff algorithm only applies in single Wireless LAN cell scenario, and the Vertical Handoff algorithm containing many Wireless LAN cell is left as upcoming work.

In [8], a new approach on vertical handoff is proposed in order to offer interoperability amongst available systems. This methodology involves a process based on loose coupling internetworking in combination with Mobile Internet Protocol version-4 under MIH (Media Independent Handoff) to provide the mobile users with seamless best connectivity in different handoff cases like imperative and alternative. A decision algorithm is proposed to implement the decision mechanism involving two access network selection functions. The proposed methodology achieves the continuation of communication session when the Mobile User Handoff among distinctive technologies available and decrease the delay and the packet loss.

Shailaja Sasi *et al*. [9] proposed a vertical handoff decision algorithm which considers the user-based preferences and network-based preferences to enable the user with guaranteed service continuity throughout a communication amongst diverse wireless networks. The model of hybrid system basically involves three stages: System discovery method being GPS-enabled, where user equipment (UE) is incorporated with a GPS technology to determine the motion orientation and UE's instant velocity at any given point of time; Handoff decision phase, where the decision to remain in the same network or to connect to another network is made; Finally vertical handover decision algorithm is utilized to find the best suitable target network technology for handoff as the UE motion direction of base stations is assessed.

III. Model Formulation

The concept of Reinforcement Learning has been used to model the environment as MDP so as to perform selection of a best network by minimizing any delay that might be encountered during the handoff decision phase.

Markov Decision Process (MDP):

Fig. 2 depicts the next generation wireless network where a Mobile Terminal is often situated in the coverage area of heterogeneous wireless networks consisting of wireless cellular network and several WLANs. In different service sector, the Mobile Terminal can access to several types of candidate networks. Hence, it is of high importance to make sure that mobile terminal can work in an optimal way without experiencing the handoff delay unnecessarily caused during the decision phase is the main problem in this paper. The Media Independent Handover Function which was recommended by IEEE 802.21 MIH Working Group [11] has a service known as Media Independent Information Service which gives information required for the link reward function, the signaling cost function parameters and network conditions estimation in heterogeneous wireless networks. The mobile terminal is able to periodically get the required information from the MIHF and choose whether to retain the connection with the present network or should be moved to the networks with a better QoS. Whereas, required maximum and minimum values or

threshold values and the value of 'ω', are defined earlier for the application considered.

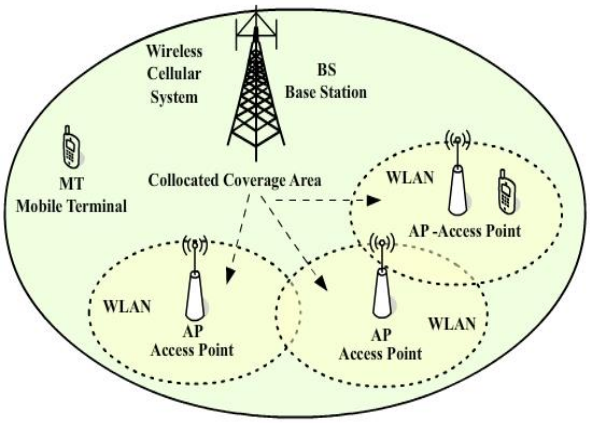

Fig. 2. Collocated heterogeneous wireless networks.

The environment for Reinforcement learning is formally defined trough the MDP where in MDP is based on a Markov property. The vertical handoff decision problem is developed or formulated as a MDP. As per this property, given the present, the future does not depend on the past. The state captures all relevant information from the history [2].

$$\mathrm{P}[S_{t+1}|S_t] = \mathrm{P}[S_{t+1}|S_1, S_2 \ldots\ldots, S_t] \quad (1)$$

The aim is to define the policy that maximizes the expected total reward per connection. Reward is a scalar value that is got for being in a state. The reward that is got by entering a state gives the usefulness of that state r(s). Where, r(s, a) gives the reward for being in a state's' and taking an action 'a' and r(s, a, s') gives the reward for being in a state, taking an action and moving in another state. The optimal policy solution for an MDP is a policy which determines the best action for each state in the MDP.

An MDP model includes five main components, namely: states, actions, decision epochs, rewards and transition probabilities.

- States: State's' consists of the facts of the network to which the Mobile Terminal is presently connected to and the available bandwidth and the average delay offered by all the available collocated networks in the area. All the parameters relating to network connectivity can be included.
- Actions: At each decision epoch t, the Mobile Terminal must decide on the handoff action a_t which is based on the current state s_t of the Mobile Terminal where $s_t \in$ S.
- Decision epochs: $[T = 1, 2 \ldots\ldots, N]$ is the sequence which represents the times of consecutive decision epochs and the Mobile Terminal has to make a decision whenever a certain time period has elapsed. N is the random variable denoting the time of connection termination and τ being the period of decision-making. (See Fig. 3)

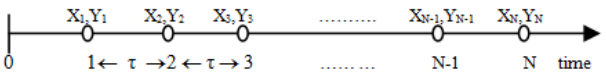

Fig. 3. Timing of Markov Decision Process (MDP) [2].

- Rewards: When any action a_t is taken, the system receives a reward $r(a_t, s_t)$ for that period.
- Transition probabilities: Transition probability gives the probability of the MT to be transitioned to next state and is given by $\mathrm{P}[s'|s, a]$.

Let the number of networks that are collocated in the coverage area of interest be denoted by M, A = {1, 2..., M} is the action set and Y_t denotes the selected action at different decision epochs. Based on the information of its current state, the Mobile Terminal selects an action. S denotes the state subspace. For each state s ∈ S, the state data consists of mobile terminal's currently connected network's identification number or address, the bandwidth availability of all the collocated networks available in the area, and the average delay provided by them. At decision epoch t, X_t is the random variable used to denote the state. For next state s′, the state transition probability function is given by $\mathrm{P}[s'|s, a]$, given current state 's' and selected action a. Subsequently the state transition relies on the present state and action and not on the previous states. This function is said to be Markovian.

The QoS received by the mobile connection within the time period (t, t +1) from the chosen network is reflected by the link reward function, given by $f(X_t, Y_t)$. Functions $\delta_t: S \rightarrow A$ is the Deterministic Markovian decision rule describing the action option, given the scheme being in state 's' at decision point t. At all the decision epochs, decision rule's sequence is used which is known as a Policy, denoted by π = (δ1, δ2..., δN). The expected total reward is calculated between all the decision epochs starting from the connection initiation until the connection terminates, where the policy π is followed for an initial state 's' [2] is given by,

$$v^{\pi}(s) = E_s^{\pi}\left[E_N \sum_{t=1}^{N} r(X_t, Y_t)\right] \quad (2)$$

where, E_s^{π} =Expectation with respect to initial state's' and policy π.

E_N = Expectation with respect to N

N= random variable denoting the connection termination time

It may be noted that different policy π and initial state 's' will change the chosen action 'a' which results in change in state transition probability function $P[s'|s, a]$ that is used in the expectation E_s^{π}. The connection termination N is assumed to have geometrical distribution with mean $\frac{1}{1-\lambda}$. Based on this, equation (4) can be rewritten as

$$v^{\pi}(s) = E_s^{\pi}\left\{\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right\} \quad (3)$$

where, λ is the discount factor of the model, 0≤ λ<1.

IV. Design and Implementation

Solving reinforcement involves three steps: formulating the environment into Markov decision process model, computation of reward function, solving the Bellman's equation using Value iteration algorithm to estimate the total expected reward per connection and reducing unnecessary handoffs.

Markov decision process model

Assuming that the heterogeneous networks coverage area consists of 'M' wireless networks where the Mobile Terminal has access to more than one network at any given point of time.

1. State: State 's' consists of the network information of the network to which the Mobile Terminal is presently associated to or base station identification number and the available average delay and bandwidth etc. Entire parameters relating to network connectivity can be included. The mobile terminal is presumed to intermittently receive information from the collocated networks within its acceptance range. The publicized information from each network may comprise, amongst other parameters, the average delay and the available bandwidth [2].

State space is given by,

$$S = N \times B \times D = (N_1, \cdots; N_M, B_1, \cdots; B_M, D_1, \cdots; D_M)$$

where, $N = (N_1 \ldots N_M)$: The number of available collocated networks.

$B = (B_1, \cdots; B_M)$: Indicates the available bandwidth provided by each of the available networks.

$D = (D_1 \ldots D_M)$: Denotes the delay provided by each of the available networks.

For simplicity, this information is assumed to be provided in multiple of units i.e. unit delay and unit bandwidth.

2. Action: The Mobile Terminal takes action based on whether the handoff is necessary or not. If yes, which is the best available network that provides low cost and high QoS. Thus, the action is given by S,

$$a = (a1, a2)$$

where, a1: Handoff is necessary

a2: Connection can be continued using the existing network

3. State transition probability: If 's' is the present state and 'a' is the action taken, the probability that s′ will be the subsequent state is given by the probability function,

$$P(s'|s,a) = \begin{cases} P[v'|v] \prod_{m \in M} P[b'_m, d'_m | b_m, d_m], & h' = a \\ 0, & h' \neq a \end{cases} \quad (4)$$

where,

s = [h, b_1, ⋯; b_M, d_1, ..., d_M, v]: Represents the current state where m=1 to M

s′ = [h′, b'_1, ⋯, b'_M, d'_1… d'_M, v′]: Represents next state

P [b'_m, d'_m |b_m, d_m]: Represents transition probability of m network's bandwidth and delay.

4. The CBR voice traffic data recommended by the ITU is used for performance evaluation using the user data gram protocol (UDP) as a transport protocol.

Reward function computation

When the Mobile Terminal is in a's' state and takes an 'a' action, then gets a reward r (s, a) immediately which can be explained as below [3]. The bandwidth reward function for the given total available bandwidth β is given by,

$$f_B(s,a) = f(\beta) = \begin{cases} 1, & \beta \geq UB \\ (\beta - LB)/(UB - LB), & LB < \beta < U \\ 0, & \beta \leq LB \end{cases} \quad (5)$$

where β: indicates the total available bandwidth

LB: Indicates the minimum required bandwidth by the connection

UB: Indicates the maximum required bandwidth by the connection

The delay reward function 'τ' represents the maximum delay is given by,

$$f_D(s,a) = f(\tau) = \begin{cases} 1, & 0 < \tau \leq LD \\ (UD - \tau)/(UD - LD), & LD < \tau < UD \\ 0, & \tau \geq UD \end{cases} \quad (6)$$

where,

$$\tau = \max\{d_i \cdot a_i\}$$

d_i indicates the network delay,

i indicates the network i = 1, ⋯M

LD and UD indicates the minimum and maximum delay required by the connection

The handoff cost function is given by

$$q(s,a) = \begin{cases} K_{h,l} & h \neq l \\ 0, & h = l \end{cases} \quad (7)$$

$K_{h,l}$: denotes the handoff cost imposed while switching from network 'h' to network 'l'.

Therefore, given the present state s and the chosen action a the reward function r (s, a) is given by,

$$f(s,a) = w f_B(s,a) + (1-w) f_D(s,a) \quad (8)$$

$$r(s,a) = f(s,a) - q(s,a) \quad (9)$$

where, w: denotes the weight factor $0 \leq w \leq 1$

Bellman's optimality equation calculation

The optimal state value function is the maximum function over all polices. The expected total reward and the optimal policy are determined using the Value Iteration Algorithm (VIA). The best network to be selected given the current state's' is indicated by the

optimal policy δ*(s). Value iteration is used for the computation of the optimal state value function by iteratively improving the estimate of V(s).

So, the description can be broken down to ensure an optimal policy in terms of finding the optimal policy from all states the Mobile Terminal can end up in. Once that is done, next step is to do the one step look ahead (Fig. 4) and discover what is the best first action that could have been taken. This is the principle of optimality applied to policies. If the solution to the sub problem V*(s′) is known, i.e. the optimal value function from s′, then V*(s) can be found by one step look ahead [10]

$$v_*(s) \leftarrow \max_{a \in A} R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \tag{10}$$

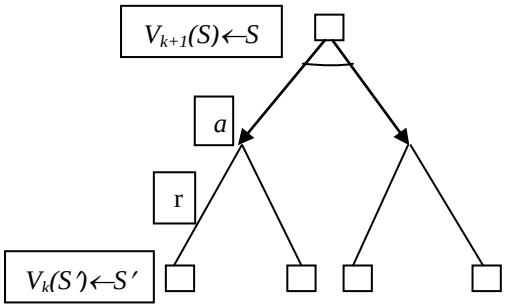

Fig 4. Value function one step look ahead [10]

In value iteration, Bellman optimality equation is iteratively applied to obtain the optimum value function. At each iteration k+1 update $V_{k+1}(s)$ from V_k (s′) for all state 's' ∈ S.

$$v_{k+1}(s) = \max_{a \in A} \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s') \right) \tag{11}$$

$$v_{k+1} = \max_{a \in A} R_s^a + \gamma P^a v_k \tag{12}$$

At every iteration, each state gets a turn to be the root. It starts off by an old value function V_k (s′) and is put in the leaf. Bellman optimality equation is taken and is turned into an iterative update. This algorithm is shown (Fig. 5) as a pseudo-code in the following [13]:

Initialize V(s) to arbitrary values
  Repeat
    For all $s \in S$
      For all $a \in A$
$Q(s,a) \leftarrow E[r|s,a] + \gamma \sum_{s' \in S} P(s'|s,a)V(s')$
$V(s) \leftarrow \max_a Q(s,a)$
Until V(s) converge

Fig. 5. Value iteration algorithm pseudo-code [13]

Let v(s) represent the maximum expected total reward, given the initial state 's'. That is,

$$v(s) = \max_{\pi \in \Pi} v^{\pi}(s) \tag{13}$$

The optimality equation or the value function is given by,

$$v(s) = \max_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} \lambda P[s'|s,a]\, v(s') \right\} \tag{14}$$

V. Results and Discussion

The main objective of developing a novel vertical handoff algorithm that can perform selection of network based on context information e.g. quality-of-services parameters like Bandwidth and delay. The complete implementation has been carried out using Matlab. This work essentially concentrated on the handoff decision phase in direction to choose the best network with maximum bandwidth and minimum delay during the handoff period. The delay encountered in the decision phase has been achieved by minimizing the number of handoffs. In the course of this assessment, it was also verified for the number of vertical handoffs and expected total reward per link as the performance parameters (Fig. 6 and Fig. 7).

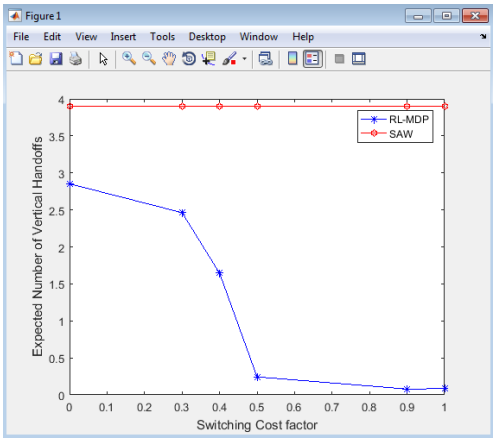

Fig. 6. Comparative analysis on total number of handoffs under different switching costs.

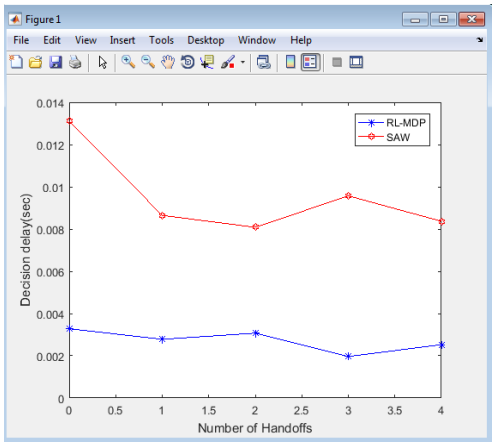

Fig. 7. Comparative analysis on number of handoffs and decision delay.

Well the study also considers comparing the outcomes with that of Simple Additive Weighting method (SAW) used to choose the best network for the continuous connection by the mobile terminal. The performance of the proposed algorithm is assessed under voice traffic and

the numerical results indicate good performance improvement of proposed scheme over the SAW technique.

A scenario where the MT is located in an area that has WLAN and Cellular network connectivity is considered. A mobile terminal must connect to at least one network during its transmission time. Agent's performance can be measured by two parameters: The expected total rewards or the value function and the expected total count of handoffs. These parameters will be with respect to per connection. The value function is given by equation 16. Notation discussed in methodology is followed throughout the implementation. For each connection, the expected count of vertical handoffs [2], is given by

$$\varsigma^{\delta}(s) = E_s^{\delta}\left\{\sum_{t=1}^{\infty} \lambda^{t-1}.1[a_t \neq i(t)]\right\} \quad (15)$$

where, $1[a_t \neq i(t)]$: represents an indicator function

i.e., $1[a_t \neq i(t)] = 1$, if a_t for time $t \neq i(t)$

0, otherwise

The time unit which is the time between the consecutive decision periods is assumed to be 15s. A heterogeneous system environment with two networks being collocated (that is M =2). Where, the first network is assumed to be a WLAN, and the second network is assumed to be a wireless cellular system. The simulation metrics or parameters of the two networks used in the simulation and the numerical results are tabulated in Table I. For convenience of calculation, moving or switching costs of the two networks are assumed to be the same (that is $K_{1,2} = K_{2,1}$).

TABLE I: NETWORK PARAMETERS USED IN SIMULATION [2]

Parameters	Network_1 (WLAN)	Network_2 (Cellular)
Maximum bandwidth in network (Units)	25	10
Maximum delay in network (Units)	8	8
Network Switching cost	1 (from 1 to 2)	-
Network Switching cost	-	1 (from 2 to 1)

TABLE II: MAXIMUM AND MINIMUM SIMULATION PARAMETERS FOR REWARD FUNCTION CALCULATION [2]

Parameters	Values
LB (Units)	2
UB (Units)	4
LD (Units)	2
UD (Units)	7

The voice traffic data i.e., Constant bit rate data is used for simulation inputs. The parameters used in the simulation to calculate the reward function [2] are tabulated in Table II.

For ease of calculation, the bandwidth and delay parameters are taken in terms of units where one bandwidth unit equals 48 kb/s and delay of 60ms. The values of LB and UB are taken in order to match with the protocols and voice coders corresponding to the multimedia services of IP. The LD and UD are taken in order to meet the target delay which is required to be less than 150ms. The acceptable connection should have delay which is in between 150ms and 400ms. The connection quality is considered as unacceptable when the delay goes beyond 400ms. All these values are taken according to the ranges recommended by ITU (International Telecommunication Union) [2].

For the network's state transition probability function of cellular system, the bandwidth and delay values are considered to be assured for the length of the connection where as for WLAN they are changing with the current traffic.

For the Bellman equation calculation, ε is considered to be 10^{-3}. The expected count of handoffs is calculated using the Value iteration algorithm (VIA).

The average time of connection length is assumed to be 10 minutes i.e., represented by $\lambda = 0.975$. $K_{i,a}$ is used to represent switching cost factor which is provided by the network operators. This value of switching cost is used to represent the complication involved around handoff process on the network while rerouting. If the $K_{i,a}$ value is low, it means that there is agreement between the networks on interworking and roaming which simplifies the process of vertical handoff. The higher values indicate that there is no agreement and the network switching cost is very high between the two networks.

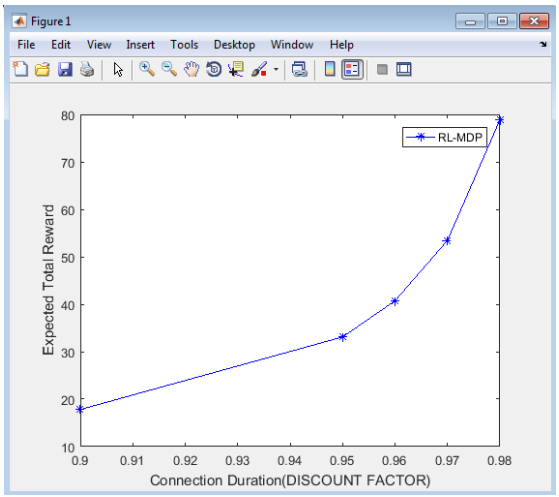

Fig. 8. Effect on expected total rewards under different connection durations (discount factor).

The Fig. 8 above shows that the expected total reward per connection increases with the increase in the link time. The agent learns with every iteration to take the right handoff decision and tries to improve its performance with every iteration. So, as the link duration increases the number of iterations also increase and the agent gets more opportunities to learn and improve its performance. Hence, the expected total reward increases with increase

in the connection duration. The time unit which is the time between the consecutive decision periods is assumed to be 15s that is ¼ min. where, average connection period being equivalent to 1/(1-λ) times the unit. The variation of λ from 0.9 to 0.98 resembles the variation of average length of connection from 2.5 minutes towards 15 minutes.

The Table III shows the expected total reward for different combination of bandwidth and delay offered by various networks which helps in selection of best network. The value of discount factor λ varies from 0.9 to 0.98 and weight factor 0.25. As per the table network 5 has maximum bandwidth 25 units and minimum delay 2 units and the total expected reward we get in the range of 17.8169-78.8624 which is highest when compared to other combinations. Hence network 5 can be considered as a best network amongst all other networks so that with minimum delay network can be selected.

TABLE III: EXPECTED TOTAL REWARD OFFERED BY DIFFERENT NETWORKS UNDER VARIOUS BANDWIDTH AND DELAY CONDITIONS

Network	Bandwidth (units)	Dealy (units)	Expected Total reward
1	5	8	2.4596-12.2580
2	15	8	6.1477-25.9681
3	20	8	8.0090-32.8405
4	25	8	9.8702-39.7130
5 [Best network]	**25**	**2**	**17.8169-78.8624**
6	25	6	11.8169-48.8624
7	10	2	12.2332-58.2450
8	10	6	6.2332-28.2450

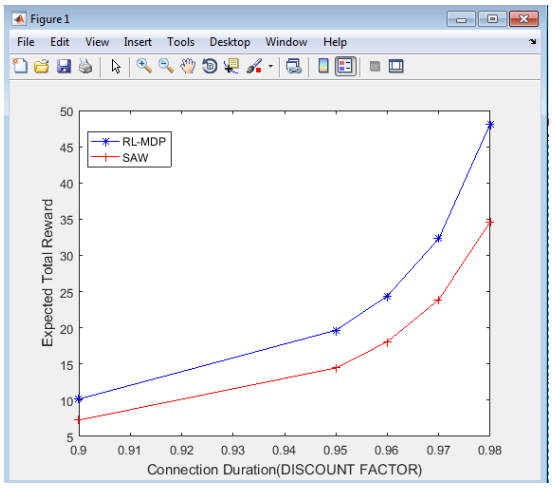

Fig. 9. Comparative analysis of expected total rewards under different connection durations (discount factor).

Fig. 9 shows the comparative analysis of expected total rewards under different connection durations (discount factor) for the proposed Reinforcement Learning based on MDP (RL-MDP) and Simple Additive Weighting technique (SAW). The RL-MDP algorithm gives the highest expected total reward for all values of λ varying from 0.9 to 0.98.

TABLE IV: EXPECTED TOTAL REWARD AND OPTIMAL POLICY

Discount Factor	No. of Iterations (Epsilon criteria satisfied)	Final value		Optimal value (Expected Total rewards)	Optimal Policy
		Action 1	Action 2		
0.98	266	44.9960	44.8461	44.9960	1
		47.4969	47.4219	47.4969	1
		45.1490	45.2460	45.2460	-1
		47.5734	47.7469	47.7469	-1
		45.0501	45.3991	45.3991	-1
		47.5240	47.9485	47.9485	-1
		44.9491	45.5502	45.5502	-1
		47.4735	48.1490	48.1490	-1

Table IV gives optimal policy based on expected total reward for λ=0.98, ε =0.001 and w=0.25. Initially state values are found out for all given states for both the actions until it satisfies the epsilon criteria and then find out the optimal value and the optimal policy. Optimal policy helps in finding the maximum expected reward of states and also to reduce the unnecessary handoffs eventually leading to reduction in handoff delay.

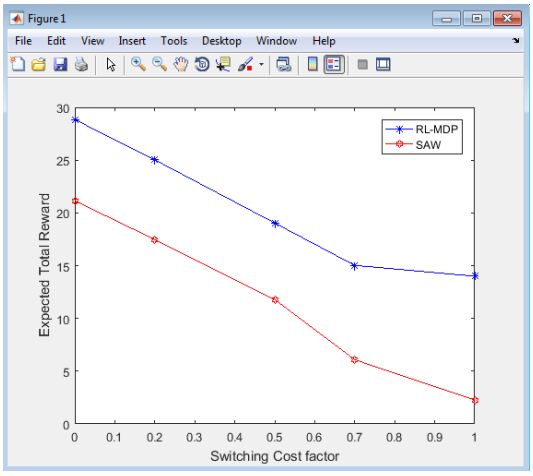

Fig. 10. Comparative analysis of expected total rewards under various switching costs

Fig. 10 shows the variation of expected total reward with respect to the switching cost. Here, the average

linking duration is taken as λ = 0.975 which equals to 10 min. The plot shows that when switching cost increases, the agent knows this is a negative factor for the performance and thus, the expected total rewards decrease with increase in the switching cost.

TABLE V: EXPECTED TOTAL REWARD AND SWITCHING COST FOR RL-MDP AND SAW

Switching Cost	RL-MDP	SAW
	Expected No. of Vertical Handoffs	Expected No. of Vertical Handoffs
0	28.8450	21.1365
0.2	25.0227	17.4729
0.5	19.0357	11.7857
0.7	15.0239	6.0857
1	14.0000	2.2739

Table V gives the Expected Total reward and switching cost for RL-MDP and SAW for λ = 0.975 and w=0.25. In both cases as switching cost increases, expected total reward decreases.

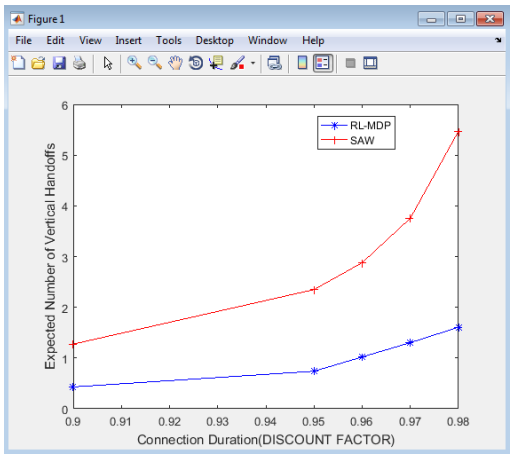

Fig. 11. Effect on total number of handoffs under different connection durations (discount factor).

TABLE VI: EXPECTED NO. OF VERTICAL HANDOFFS AND DISCOUNT FACTOR FOR RL-MDP AND SAW

Discount factor	RL-MDP	SAW
	Expected No. of Vertical Handoffs	Expected No. of Vertical Handoffs
0.9	0.4299	1.2741
0.95	0.7396	2.3491
0.96	1.0203	2.8791
0.97	1.3036	3.7564
0.98	1.6059	5.4725

Fig. 11 plots the expected count of handoffs with respect to the connection duration. The plot shows that the total number of vertical handoffs increases with the increase in the connection period. When there is an increase in the average duration of connection, the decision epoch will increase in number. Therefore, the expected number of vertical handoffs increases for SAW but RL-based MDP algorithm has quite steady number of vertical handoffs for its static policy. The reduction in decision delay has been achieved by minimizing the number of unnecessary handoffs. Table VI also depicts the same with the numerical values comparison between RL-MDP and SAW.

Fig. 6 shows the expected number of vertical handoffs with respect to the network switching cost. It is seen that when there is an increase in the value of switching cost factor, there is less encouragement to take action on performing vertical handoff because the rewards will be decreased. So, the agent chooses not to go for the handoff and avoids switching more frequently. Therefore, the count of total vertical handoffs comes down with the increase in the switching cost. When the switching cost increases, the RL-MDP algorithm achieves the less vertical handoffs than SAW algorithm. While the SAW does not take switching cost into consideration, the expected number of handoffs remains unaffected. Table VII gives numerical values comparison of Expected No. of Vertical Handoffs and switching costs for RL-MDP and SAW.

TABLE VII: EXPECTED NO. OF VERTICAL HANDOFFS AND SWITCHING COSTS FOR RL-MDP AND SAW

Switching Cost	RL-MDP	SAW
	Expected No. of Vertical Handoffs	Expected No. of Vertical Handoffs
0	2.8527	3.9000
0.3	2.4629	3.9000
0.4	1.6447	3.9000
0.5	0.2430	3.9000
0.9	0.0787	3.9000

A 5% reduction in decision time to handoff when appropriate signaling cost, bandwidth and delay are considered has been achieved in our proposed work. Assuming that a user encounters 4 handoffs during his connectivity period, it is seen from Fig. 7, a 5% reduction for decision to handoff has been achieved as compared to the decision technique used in Simple Additive Weighting method (SAW) algorithm.

VI. CONCLUSION

Reinforcement learning in the handoff decision phase has been used to make appropriate decisions during the decision phase of vertical handoff of wireless heterogeneous networks. MDP is used to model the environment. The aim is to increase a connection's expected total reward. The performance of the agent is evaluated using a reward function which considers the QoS parameters and the network switching cost of the mobile connection. An optimal policy that results in optimal value is obtained in the results after several iterations which give the maximum expected total rewards per connection which helps in selection of best network with minimum delay. The voice traffic data recommended by the ITU is used for performance

evaluation. The results indicate that the performance of the proposed algorithm improves as the number of iterations increases with the increase in connection duration and it gets more opportunities to learn and take better decisions actions. The results show that by taking proper decision for handoff, we can efficaciously reduce unnecessary handoff's that leads to increase in traffic load that in turn might lead to packet loss and call blocking. The proposed system also tries to avoid unnecessary handoffs when the network switching cost is high making the solution a cost effective one. A reduction in decision delay along with the minimum number of handoffs has been achieved. The overall performance of the proposed algorithm is assessed under voice traffic and the numerical results show good performance improvement of proposed scheme over the SAW technique. The overall performance of the proposed algorithm is evaluated beneath voice site visitors and the numerical results show the appropriate overall performance improvement of the proposed scheme over the SAW approach.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Both authors conducted the research, analyzed the data; wrote the paper; and had approved the final version.

References

[1] S. Akhila, J. K. Murthy, A. R. Shankar, and S. Kumar, "An overview on decision techniques for vertical handoffs across wireless heterogeneous networks," *International Journal of Scientific & Engineering Research*, vol. 3, no. 1, January 2012.

[2] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, 2008.

[3] R. Chai, W. G. Zhou, Q. B. Chen, and L. Tang, "A survey on vertical handoff decision for heterogeneous wireless networks," in *Proc. IEEE Youth Conference on Information, Computing and Telecommunication*, 2009, pp. 279-282.

[4] Y. Chen, H. Chen, and L. Xie, "An MDP-based handoff decision algorithm for multi-domain heterogeneous wireless access networks," in *Proc. International Conference on Communications, Circuits and Systems (ICCCAS)*, 2010, pp. 163-167.

[5] M. Bin, D. Hong, X. Xianzhong, *et al*., "An optimized vertical handoff algorithm based on Markov process in vehicle heterogeneous network," *China Communications*, vol. 12, no. 12, pp. 106-116, 2015.

[6] H. Tabrizi, G. Farhadi, and J. Cioffi, "A learning-based network selection method in heterogeneous wireless systems," in *Proc. Global Telecommunications Conference (GLOBECOM 2011)*, 2011, pp. 1-5.

[7] G. Ren, J. Zhao, and H. Qu, "A user mobility pattern based vertical handoff decision algorithm in cellular-WLAN integrated networks," in *Proc. Second IEEE International Conference on Computer and Communications*, 2016, pp. 1550-1554.

[8] S. Gueziz and D. Korichi, "Performance analysis of handover optimization based on Media Independent handover in new networks NGWN," in *Proc. International Symposium on Networks, Computers and Communications (ISNCC)*, 2016, pp. 1-6.

[9] S. Sasi, R. D. Daruwala, T. Palav, and P. Mule, "A hybrid vertical handoff decision algorithm for seamless mobility in heterogeneous wireless networks," in *Proc. International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 1921-1923.

[10] David Silver. RL course by [Online]. Available: https://medium.com/biffures/rl-course-by-david-silver-lectures-1-to-4-7667608bf7d3

[11] The IEEE 802.21 Working Group finished all PAR Related Activities and Entered a State of Hibernation on July 19, 2019

[12] Markov Decision Process. [Online]. Available: https://en.wikipedia.org/wiki/Markov_decision_process

[13] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, 3rd edition, 2014.

Hemavathi is working as an Assistant Professor in the department of Electronics and Communication Engineering at B.M.S. College of Engineering, Bengaluru. She has completed her B.E in Electronics and Communication from Adi Chunchanagiri Institute of Technology, Chikmagalur, Kuvempu University, M.Tech from Acharya Institute of Technology, Bengaluru, VTU (Visvesvaraya Technological University). Currently she is pursuing Ph.D in the department of Electronics and Communication Engineering, B.M.S. College of Engineering at VTU. She is a life member of Indian Society for Technical education (ISTE) since 2018. She has published papers in 6 International Journals, 1 International Conference and 4 National conferences..

Dr. S. Akhila received her Bachelors in Electronics in the year 1988 and her Masters in Electronics in the year 1994 from University Visvesvaraya College of Engineering, Bangalore, India. She has completed her Ph.D. in the year 2013 from the Visvesvaraya Technological University (VTU) in Wireless Communication. Since 1995, she has been with B.M.S. College of Engineering, where she is working as a Professor in the Electronics and Communication Engineering Department. She is a life member of Indian Society for Technical education (ISTE). She has published papers in more than 4 National conferences, 6 International Conference and 25 International Journals.