Reinforcement Learning Algorithm for Routing in Space-Air-Ground Integrated Networks

Amare Haymanot Hailu, ID: M202261025, Major: Information and Communication Engineering

***Abstract*—Space Air Ground Integrated Network and 6 Generation (6G) communication technologies have entered the field of study by gaining intense research interest. SAGIN has evolved as a pattern shifting architecture by providing flexible or adaptable wireless coverage and seamless, large-scale, high-rate connectivity to support terrestrial communications. It is a highly heterogeneous, self-organizing, and time-varying wireless network that provides extensive and global connection, in contrast to conventional ground or satellite networks. As a result, it has significant difficulties in routing design. Mechanism of routing will be affected by the time-changing dynamic mobile network making it challenging to achieve the optimal traffic delivery performance. Therefore, the main concern of routing is the routes optimization when we still have dynamic topology variations. To address the above challenges, an integrated space-air-ground routing system based on reinforcement learning techniques and principles are discussed in this report. Reinforcement learning (RL),which is a subset of machine learning, provides a framework for systems to learn from previous interactions or experiences with the environment to efficiently choose its actions in the future. RL is appropriate and suitable for addressing optimization problems specially related to distributed systems and for routing in networks, particularly in SAGIN. Compared to other optimization techniques used to solve the same problems, reinforcement learning also has overhead—in terms of computation, memory and control packets. Reinforcement learning (RL) can efficiently resolve the problem of limited satellite bandwidth resources and hot air balloons energy in the space-air-ground integrated network, and can implement the routing function.**

***Index Terms*—6G communication, SAGIN, Routing, Reinforcement learning.**

I. Introduction

Although the communication industry is improving every time, the recent different communication networks have challenges and problems. When we see the ground communication network, it has communication weak areas, that are easily affected by geographical considerations and factors. The air communication network has limited energy and the network is not reliable [1] On the other hand, the satellite communication networks have difficulties with restricted bandwidth and long delays and limited hot air balloon energy. With the improvement of user needs, a new trend has progressively arisen the combination of space, air and ground. Space–air–ground integrated networks (SAGIN) is the integration of ground networks, satellite systems, and air networks to ensure the throughput and dependability (reliability) of data transmission [2]. The low-cost air-based network becomes the relay between the space-based and ground-based. As one of the most promising networks at present, the integration of the three components (space, air and ground) has the three characteristics of collaboration, high efficiency and ubiquity which can meet the various needs of users in various fields. But, SAGIN encounters a series challenges and difficulties due to its heterogeneous, self-organizing, time-varying, distributed, open and other behaviours. Such as routing and switching management, network control, system design, security protection, spectrum management, energy management. In this case, we focus on one of the big challenges which is routing.

Big capacity data transmission between multiple networks is needed in the integration of space, air and ground, therefore, it is important to study routing between multi-layer networks. In emergency conditions, Qu [3] suggested a load-balancing dynamic routing approach that significantly lowers the packet loss rate and delay. But OLSR has a significant routing overhead. Only for UAV ad hoc networks did Zheng [4] sequentially introduce multi-path routing techniques, mobility perception and load perception. Tan [5] suggested a QoS-OLSR routing in SAGIN scenario, but only for a cluster structure. None of the aforementioned routing plans account for the dual issues of insufficient satellite bandwidth and restricted aircraft fuel. Consequently, various researchers have applied reinforcement learning (one of a sub set of AI) to the routing field. Xiong [6] used Q-learning to study network's flexibility of nodes as a reference basis for next-hop routing, that is appropriate for mobile ad hoc network with fast node mobility. Deep reinforcement learning is applied by Xu [7] to develop an intelligent joint resource allocation method for maritime communication networks, which efficiently improves communication and computing effectiveness, but only considers static scenarios. This shows the popularity of reinforcement learning in the routing field.

II. Architecture for Space–Air–Ground Integrated Network

The architecture of a SAGIN is shows in the Figure 1 below, consisting of a space network, air network, and ground network, which enables the sharing of resources and global information by connecting several networks to shape a complex topology [8]. Each satellite has two intra-orbit links and two inter-orbit links, and there are no inter-orbit links between polar regions and reverse seams. And also, there is no connection or links between the users.

1) Space Network: The space network includes satellites and constellations and also their corresponding terrestrial infrastructures, such as ground stations and network operations control centres. These satellites and constellations have different characteristics and are in different orbits. GEO, MEO, and LEO satellites are the three subgroups of satellites [9].

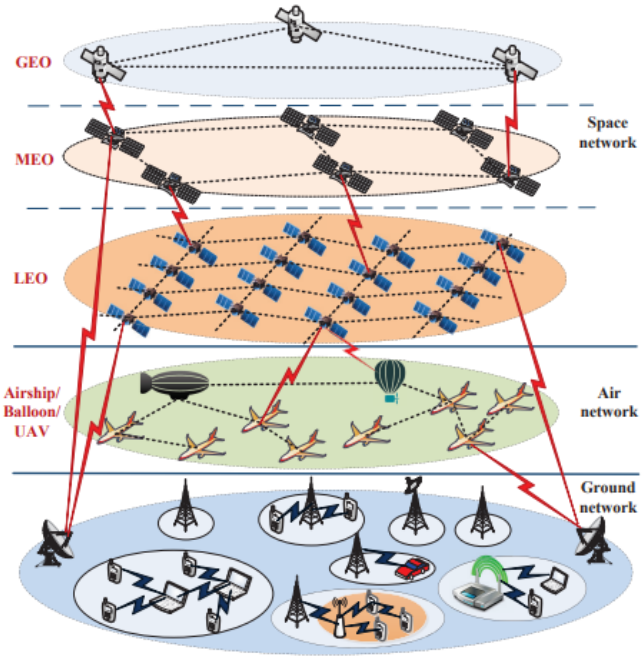

Fig. 1: Architecture of SAGIN.

Additionally, we can divide satellite networks into narrow band and broadband based on the channel bandwidth.

2) Aerial Network: The air network is an aerial mobile system that uses aircraft as carriers for information processing, transmission, acquisition. Airships, balloons and UAVs are the main infrastructures that make up the high and low altitude platforms (HAPs LAPs) which can provide broadband wireless communications complementing the terrestrial networks. Compared to base station (BSs) in terrestrial network, air network has the features of low cost, easy deployment, and large coverage to offer wireless access services on a local basis.

3) Ground Network: The ground/sea network has terrestrial communication systems like mobile ad hoc network [10], wireless local area networks, Worldwide Interoperability for Microwave access [11], cellular network, etc. Mainly, cellular network has grown from the first generation (1G), to the second generation (2G) and third generation (3G) through the fourth generation (4G) or Long-Term Evolution-Advanced (LTE-A), and nowadays it is developing to 5G wireless network, to provide numerous services. In terms of standardization, the Third Generation Partnership Project (3GPP) has established specifications for cellular networks. Even though terrestrial networks can offer users with high data rates, their coverage in remote and rural areas are quite limited.

A. Advantages of SAGIN Architecture

The architecture of Space Air Ground Integrated Network (SAGIN) allows three network layers to operate independently and work together. Combine heterogeneous networks to build layered wireless broadband networks. The satellite can efficiently reach rural and remote areas and significantly reduce the cost of fifth generation (5G) terrestrial networks. At the same time, it offers advantages in terms of coverage, throughput, reliability, and flexibility, offering a wide range of application possibilities in next-generation networks. Moreover, SAGIN is very significant in many situations such as remote sensing and navigation and communications.

III. Routing Algorithm

Routing algorithms address the problem of finding the most connected End to End (E2E) path to successfully transmit data traffic to the destinations with a reduced E2E latency. Because of the particular features of space and air segments in SAGIN, such as dynamic topology, non-homogeneous traffic distribution, limited power and processing capabilities, it is necessary to develop appropriate routing approaches to manage and optimize network resource. In an integrated space-air-ground network, different traffic flows should be routed to different space ground, air-ground, and space-air links according to their different QoS requirements. Early routing protocols in NGEO satellite systems were usually connection oriented, they assumed asynchronous transfer mode (ATM)-like switches in the satellites, and most routing algorithms in satellite systems were ATM-based [12] [13]. As Internet is becoming very popular and the efforts concerning the next generation satellite networks are on the way, there has been an initiative to use IP routing technology in satellite networks.

Nowadays, Artificial Intelligence (AI) is developing and it is stimulating numerous applications in wireless communications. Taking the routing problem as an example. Yet, the network topology was assumed to be static and the global node status had to be known for making routing decisions. To be able to handle highly dynamic network topologies, a routing algorithm supported by deep reinforcement learning (DRL) was designed for AANET. By relying only on local information rather than global information, the recommended high-performance DRL-assisted routing can achieve near-optimal end-to-end (E2E) latency.

IV. Reinforcement Learning Routing

A. State, Action and Reward

The Agent(learner), interacts with the environment and, depending on its current state and the reinforcement it receives from the environment, chooses its actions to apply to the environment (it' s shown in the Fig. 1 below). For instance, a router interacts with its neighbour nodes to apply routing decisions. In this scenario, the router represents the agent, the environment is router's neighbour and selections of next neighbour nodes to transmit data packets are actions. Algorithms of RL are based on a reward functions. The role of the reward (returned by the environment to the agent), is to deliver feedback to the learning algorithm about the result of the recent taken action. A reward function shows what is good/bad in the short term, while the value function shows what is good/bad in long-term.

A reinforcement learning problem is represented by a 4-tuple. These are, states, actions, a matrix of state transition probabilities, and a reward function. Environmental models

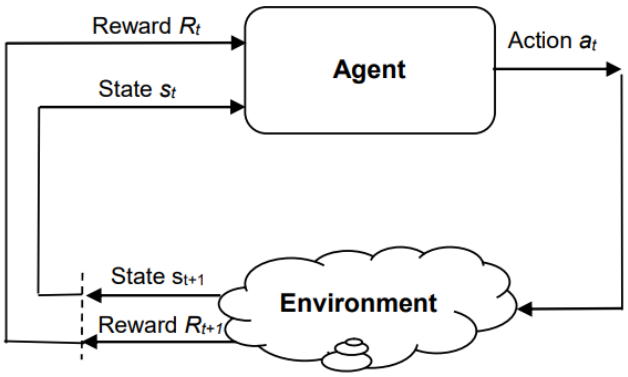

Fig. 2: Reinforcement learning model.

are described in terms of probabilities and rewards. There are two ways to solve the reinforcement learning problem: Model-based and model-free methods.

A rule t defines how the learning agent performs at time t. denotes the probability that: at=a if st=s. Where, st and at represent the state and action at time t). The probability to move, at time t, from states to state s' by taking action a is:

$$P(s,a,s') = pr(s'|s,a) - pr\{s_(s+1) - s'|s_t - s, a+t-a\} \quad (1)$$

$$\sum_{s' \in s} P(s'|s,a) = 1 \quad (2)$$

The received reward at time is a real number denoted by Rt. The reward of being in state s and taking action also is denoted by R(s,a). An action is selected by the agent chooses an action at each step (learning period) of its lifetime.

B. Q-Learning

An approach to estimate action function is proposed by Watkins. Watkins's action-function is called Q-function and the learning technique is called Q-learning. The latter is a model-free learning technique. Watkins's estimation of action function is independent of the policy followed by the agent, that makes Q-learning applicable in many circumstances and easy to implement. So, we have:

$$Q_{\pi^*}(s_t, a_t) \triangleq Q^*(s_t, a_t)$$
$$V^*(s) = \max Q(s,a), a \in A \quad (3)$$

In Q-learning, the agent learning involves epochs or number of iterations, which is a series of phases. In epoch n, the agent is in state, it performs action, it receives a reward, and it moves to state.

In reinforcement learning principle, the selection of action behaviour is optimized through a huge number of offline interactions between the agent and the environment to learn the global optimum. Once the training is completed, the routing strategy can calculate the approximate optimal solution in a short time. The calculation basis is Q-value square matrix, so it' s time complexity is only n2. Knowing the initial state of the network and the data circulation of each time slot, you can know the network status of different time slots. Its essence is the Markov decision process, which is suitable for solving decision-making problems. Our goal is to avoid energy consumption and bandwidth waste under the premise of minimizing time delay, and to solve resource scheduling problems. The environment is the entire space-air-ground integrated network, and the state space S is all nodes. Action space A is all optional links. Its input is the network adjacency matrix and the agents output the best route selection in the current state. The effective action set will change with the transfer of the data packet position, and stop learning when it reaches the destination node. The Q action value update rules are as follows.

$$Q_n(s_n, a_t) = (1-\alpha)Q_{n-1}(s_n, a_n) + \alpha[R_n + \gamma * maxQ_n - 1(s_{n+1}, a)] \quad (4)$$

where alpha is the learning rate/factor. The initial Q values, Q0 (s,a) for all states and actions are assumed given.

We can also write the updating of Q-function value in discrete time t (the state and action at time t), that is most of the time used in literature relating to Q-learning applications:

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha[R_{t+1} + \gamma^* \max Q(s_{t+1}, a)] \quad (5)$$

In order to avoid a condition where an action is never selected, the algorithm has to learn to explore new actions. So, a greedy strategy implemented. The principle is:

$$a_t = \begin{cases} argmax_a Q(s,a), 1-\epsilon; \\ random, \epsilon; \end{cases} \quad (6)$$

Where is the probability of arbitrarily chosen action and is a dynamic attenuation value to assure the complete investigation of the path. If it is visited, the action will be selected again. The design of the reward function R, which is an evaluation of various selections made in different positions is as follow:

$$\begin{cases} r_E 1 - E_{i,r} < \eta(\lambda) \\ r_B B_{rate^{i,j}} < \eta_2(\lambda) \\ 0 destination \\ -1 else \end{cases} \quad (7)$$

Where rB and rE are the penalties produced after the bandwidth utilization is very high and the remaining energy of the node is very small. They are negative numbers with relatively big absolute values. The thresholds of all parameters are a function of .

The key idea of reinforcement learning is to iterate over each learning. The single learning process is, assuming that the initial state is the source node s. If the destination node d is not reached, complete the following procedures: 1. Choose the way-out link a of the s node based on the greedy strategy. 2. Updating R (reward value). 3. Updating the Q value table based on the formula (the Q value is reduced by 100 if it exceeds the environmental boundary). 4. Updating the next state and also ready for the next cycle. Watkins stated that Q-learning converges to the optimum action-values with probability as long as all actions are repeatedly sampled in all states. Therefore, Q-learning is the most effective and efficient learning method for learning from delayed reinforcement (learning or based on reward that can be received far in the future).

C. Q-Routing Protocol

The name of the proposed algorithm comes from the notation of Q-function used in Q-learning method. The algorithm may be briefed as follows: Suppose i represent the node holding a packet P to be forwarded and where Qi (d, j) the delivery delay (E2E delay) that node i approximates it takes, for node j, to deliver packet P to the destination d. Node i preserves a table including Q-values (estimates of the transfer delay); a Q-value is associated with each neighbour of node i. When node i has a packet to send, it selects a node j with the lowest Q-value. Upon sending packet P to node j, node i immediately gets back j's an estimate for the time remaining in the trip to destination d denoted by j(d):

$$\Theta_j = minQ_j(d,k), k \in N_g(j) \tag{8}$$

Where Ng(j) is the set of j's neighbours. And then, delivery delay estimate (associated with neighbour j) is updated by node i as follow:

$$Q_i(j,d) = (1-\alpha)Q_i(d,j) + \alpha * (qt_i + T_x.T_{i,j} + \theta_j(d)) \tag{9}$$

Where alpha is the parameter of learning, qti is the time spent by packet P in i's queue and Tx T(i,j) is the transmission time between nodes i and j.

V. Reinforcement Learning Routing

Figure 3 explains the problem to be solved by reinforcement learning routing. In figure 4 below, when L1, L4, L7 send data to L6 simultaneously, if selected based on the shortest path algorithm, L5 becomes a required node, that can cause bandwidth (energy) insufficiency, packet loss and congestion. If the same condition happens again, then similar problem will occur. Reinforcement learning (RL) has a memory function, so that when path congestion occurs, it will select a new path, like through L2.

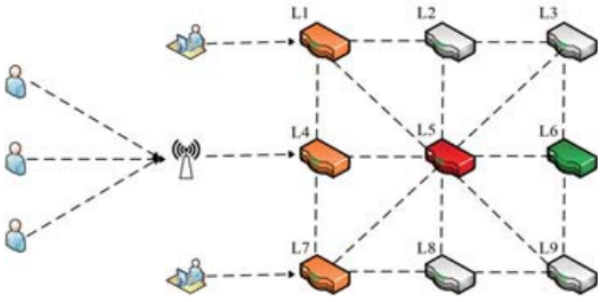

Fig. 3: Reinforcement learning model.

VI. Application of RL to Routing Protocols

A. Major Concerns and Components in RL-based Routing Design.

The following aspects can be addressed by RL-based design: 1.To identify the most appropriate states and actions of agent 2.To identify of the environment when model when possible. 3.To define a reward function that depends on metrics to optimize Various design models for the desired application field can be elaborated. The way these models solves the previously mentioned aspects are different. However, it is the same, when it comes to routing in our network. In last 25 years, different RL based routing protocols are introduced. A high-level structure of routing is shown in figure 4 below.

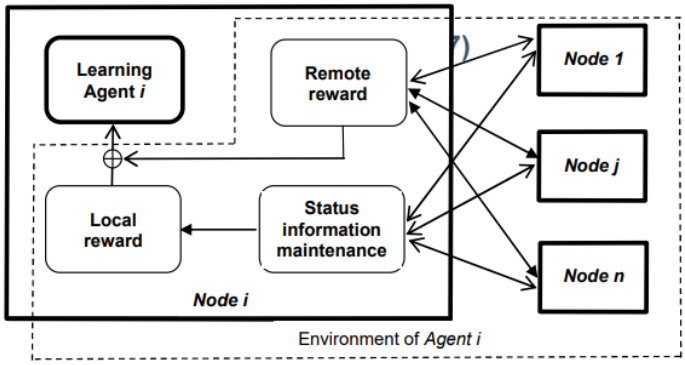

Fig. 4: High level structure of routing based on RL.

B. Network Characteristics and Classes Solved in RL-Based Protocols

Depending on their characteristics regarding mobility, the type of mediums, energy constraints, and so on, networks of communication are mainly classified into different classes. RL-based optimization of routes impacted by the characteristics of the network. In particular, Fig.4 shows the design of the components guided by the characteristics of targeted network.In general, the network's characteristics that affect the design of RL-based protocols are: topology changes, mobility of nodes, energy consumption and lifetime, distributed or centralized control, capacity and stability of links and so on. As networks evolved, the RL-based routing protocols have been improved and developed.As a result, current RL-based protocols can handle and solve almost all network characteristics.

VII. Problem Formation

A. Route formation optimization

The goal is to reduce the total delay and ensuring energy consumption and bandwidth.

$$\max_{\lambda} \sum_{i,j} \Omega \tag{10}$$

The following two ideas are the constrictions. A function of the Poisson distribution parameter is the threshold. 1.The minimum remaining energy of all nodes is not lower than the threshold, that is, from the perspective of the overall network, select a route that balances energy consumption. The principle is as follows.

$$E_{ir}(t) >= \eta_1(\lambda) = 1 - \frac{0.9\lambda}{\lambda_{max}}, i$$

2. The bandwidth utilization rate of each link does not exceed the network congestion rate, so as to more effectively

control node access channels and improve the ability to share satellite channel resources. The principle is as follows.

$$B_{rate}^{ij} = \frac{L_{pkt}.N_{pkt}}{B_{ij}} <= \eta_2(\lambda) = \frac{0.9\lambda}{\lambda_{max}}, i, j. \quad (11)$$

Where Npkt is the number of data packets per unit time.

B. Implementation process

1. Initialization: The data stream's poison distribution parameters of the data stream, clock.

2. If it arrives when the data packet is about to be sent, configure the data packet's parameters and send it right away. According to Floyd or Q-learning, the next hop is chosen. Additionally, it is required to confirm each time that no data packets have been transmitted since the last time, and if there are, to keep sending them.

3. A new node is determined to be the destination node when a data packet arrives at it. If so, determine if all data packets have been transmitted at that time before moving on to step 4. If not, update the link bandwidth, link queue, and node remaining energy and link rate of bandwidth utilization, and iterate it (loop) it in turn.

4. The final moment of the simulation time is determined if all data packets up to this point have been delivered. Otherwise, record the incomplete data packet if it is set to 5.

5. If it is the last moment of the simulation time, judge whether it is the last Poisson parameter. If it is moved to 6, else continue to convey the data packet at the subsequent moment.

6. The process ends if it is the last Poisson parameter. Else, update the clock and the next Poisson parameter , go to 2.

Algorithm 1 Q Routing

1: **Input:**
2: $Q_i(*,*)$is the matrix of Q-value of node i. Q_i can also be initialized randomly
3: **Loop:**
4: If (packet to send is ready):
5: Select next hop j with the lowest Q-value
6: packet to node j
7: Node i immediately gets back j's an estimate for the time remaining in the trip to 6: destination d (calculated by the equation (8) above) Node i updates the delivery delay estimation by using equation (9)
8: **End for**
9: Until condition termination

VIII. Conclusion

In this report, we discussed about Space Air Ground Integrated Network(SAGIN), its architecture and Routing in SAGIN. By discussing the techniques, principles and algorithm reinforcement learning routing, we analyzed that how the it can solve the problem of limited satellite bandwidth resources and hot air balloons energy in the space-air-ground integrated network, and can implement the routing function. In this analysis, We could be able to get some insights about this field and be able to understand that the challenges of routing in SAGIN can be addressed using AI technology, Reinforcement learning. At at the limited energy and bandwidth of several nodes and congestion of SAGIN, reinforcement learning (RL) method for a routing system is discussed. The principle of routing optimization is to attain the lowest delay under the premise of ensuring remaining energy and bandwidth usage, which is attained through the design of the reward function of reinforcement learning algorithms. Efficient integrated network routing in space, air and ground can be obtained according to the different real-time processing needs of the nodes. In general, we analyzed the algorithms and methods to solve this problem, and in the future work, we will show the practical work.

Acknowledgement

I would like to thank my teacher professor Du for her great support for us to understand and study about this interesting field, 6G communication and Space Air Ground Integration Network.

References

[1] H. Dai, H. Bian, C. Li and B. Wang, "UAV-Aided Wireless Communication Design with Energy Constraint in Space-Air-Ground Integrated Green IoT Networks," IEEE Access, vol. 8, pp. 86251-86261, 2020

[2] Dai, C.; Li, X.; Chen, Q. Intelligent Coordinated Task Scheduling in Space-Air-Ground Integrated Network. In Proceedings of the 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 23–25 October 2019; pp. 1–6.

[3] H. Qu, Y. Luo, J. Zhao and Z. Luan, "An LBMRE-OLSR Routing Algorithm under the Emergency Scenarios in the Space-Air-Ground Integrated Networks," in Proc. 2020 Information Communication Technologies Conference (ICTC), 2020, pp. 103- 107.

[4] Y. Zheng, "Research on the Routing Algorithm of UAV Self-organizing Network," Dissertation, University of Electronic Science and Technology, 2014.

[5] L. Tan and T. Zhang, "Clustering-based QoS Routing Protocol for Integration Network of Near Space, Air and Ground," in Proc. 2019 IEEE 5th International Conference on Computer and Communications (ICCC), 2019, pp. 370-376.

[6] K. Xiong, X. Jin, Q. Liu, "A mobile ad hoc network routing protocol optimized based on Q-learning," Journal of Beijing Jiaotong University, vol. 44, no. 2, pp. 66-73, 2020.

[7] F. Xu, F. Yang, C. Zhao and S. Wu, "Deep reinforcement learning based joint edge resource management in maritime network," in China Communications, vol. 17, no. 5, pp. 211-222, May 2020.

[8] Liu, Jiajia, et al. "Space-air-ground integrated network: A survey." IEEE Communications Surveys Tutorials 20.4 (2018): 2714-2741.

[9] S. Chandrasekharan, K. Gomez, A. Al-Hourani et al., "Designing and implementing future aerial communication

networks," IEEE Communication. Mag., vol. 54, no. 5, pp. 26–34, May 2016.

[10] M. Conti and S. Giordano, "Mobile Ad Hoc networking: milestones, challenges, and new research directions," IEEE Communication. Mag., vol. 52, no. 1, pp. 85–96, Jan. 2014.

[11] F. Aalamifar, L. Lampe, S. Bavarian, and E. Crozier, "WiMAX technology in smart distribution networks: architecture, modeling, and applications," in Proc. IEEE PES TD, Apr. 2014, pp. 1–5.

[12] M. Werner, C. Delucchi, H.-J. Vögel et al., "ATM-based routing in LEO/MEO satellite networks with intersatellite links," IEEE J. Sel. Areas Commun., vol. 15, no. 1, pp. 69–82, Jan. 1997.

[13] M. W. and, "A dynamic routing concept for ATM-based satellite personal communication networks," IEEE J. Sel. Areas Commun., vol. 15, no. 8, pp. 1636–1648, Aug. 1997.

[14] C. J. Watkins, "Learning from delayed rewards," PhD Thesis, University of Cambridge, UK, 1989.