# Energy Efficient Bandwidth Allocation and Routing in Electromagnetic Nano-Networks via Reinforcement Learning

Mohammed A. Alshorbaji[1,2] Ahmed Lawey[1] Syed A. R. Zaidi[1]
[1]School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK
[2]Electrical Engineering Department, University of Mosul, Mosul, Iraq
Corresponding Author: Mohammed A. Alshorbaji (ml14mm@leeds.ac.uk; ahmedm86@uomosul.edu.iq)

***Abstract*— Electromagnetic nano-networks operating in the THz band offer a promising solution for enabling communication among many nanoscale devices. However, the inherent limitations of nano-nodes, such as restricted energy and processing resources and short communication range, pose significant challenges for efficient data transmission. While prior research has explored Reinforcement Learning (RL) for optimising traffic routing in electromagnetic nano-networks, this paper proposes a novel approach that jointly optimises routing and sub-channel bandwidth allocation to minimise network energy consumption using RL. We leverage the Q-learning algorithm to develop a dynamic single-hop or multi-hop routing scheme that considers each node's location, energy storage capability, and the available sub-channel bandwidth. Our model formulates a reward function that balances these multiple objectives and enables the selection of optimal transmission policies for each nano-node. Our findings suggest carefully choosing the number of hops and increasing bandwidth in sub-channels can lead to substantial energy savings in nano-networks.**

***Keywords*— *Electromagnetic Nano-Networks, Reinforcement Learning, Q-Learning, Multi-hop Routing, Bandwidth Allocation, Energy Efficiency.***

## I. INTRODUCTION

### A. Motivation

Electromagnetic nano-networks operating in the Terahertz (THz) band offer a promising communication paradigm for future applications due to their high channel bandwidth and miniaturisation potential 11. However, realising the full potential of THz nano-networks requires overcoming several key challenges: (1) the high path loss [1] and molecular absorption experienced by THz signals significantly restrict the communication range of individual nano-nodes [2]. This necessitates multi-hop communication [3], [4], where data packets are relayed through multiple nano-nodes to reach their destination. (2) the small size of nano-nodes limits their battery capacity [5], [6], hence the need for energy-efficient routing protocols that can minimise energy consumption during data transmission [7], [8]. Finally, nano-nodes have limited processing power and memory [5], restricting the complexity of routing algorithms that can be implemented on these devices. These challenges necessitate the development of novel routing protocols specifically designed for the unique characteristics of THz nano-networks. Traditional routing algorithms used in conventional wireless networks are often unsuitable [9] due to their high energy consumption and computational complexity [10]. On the other hand, a dense deployment of nano-nodes can result in significant redundancy and conflicts [11] within the nano-network. This paper focuses on addressing these challenges by leveraging the power of RL to design intelligent and energy-efficient routing protocols for THz nano-networks.

### B. Related Works

Several routing protocols have been proposed for nano-networks to optimise energy usage and extend the network's lifespan . While simple and robust, flooding-based approaches lead to high energy consumption and are not well-suited for resource-constrained nano-nodes [12]. To address this, researchers have explored ways to restrict the flooding area, leading to protocols like Coordinate and Routing System for Nano-network Algorithm (CORONA) [11], Deployable Routing system (DEROUS) [13], and Stateless Linear Routing (SLR) [14]. However, these approaches still suffer from high energy consumption and require specific network structures for assigning coordinates [15]. Additionally, they do not explicitly address the limited memory capacity of nano-nodes [9].

Single-path routing protocols offer energy efficiency for wireless nano-networks (WNNs) but can suffer from increased packet loss. To address this trade-off, protocols such as Multi-hop Transmission Decision (MHTD) [3], Energy Efficient Multi-hop Routing (EEMR) [4], Energy Conserving Routing (ECR) [16], and Time-To-Live (TTL)-based Efficient Forwarding (TEForward) [17] focus on finding the optimal transmission path that minimises energy consumption while mitigating packet loss issues.

Reinforcement learning (RL) offers a promising approach for designing adaptive and intelligent routing protocols for nano-networks. In particular, Q-learning has been successfully applied to address network congestion and optimise resource allocation in traditional networks [18]. However, adapting these techniques to THz nano-networks' unique constraints and dynamics remains an open challenge. In [12], a multi-hop deflection routing (MDR-RL) algorithm based on reinforcement learning is proposed to address the unique challenges of nano-networks, including limited transmission range, energy fluctuations due to harvesting, and constrained memory. MDR-RL utilises routing and deflection tables to dynamically explore efficient paths, considering factors like energy status, hop count, and packet loss. Overall,

they presets a new method to address nano-network memory limitations by using deflection routing driven by reinforcement learning.

In [19], reinforcement learning-based routing (MDR-RL) is designed for THz flow-guided nano-sensor networks and maximises throughput by dynamically adapting to network conditions. Simulations show that multi-hop routing, especially the two-hop configuration, significantly improves performance compared to direct communication. To the best of our knowledge, no previous studies have studied the effect of using different sub-channel bandwidths on the total nano-network energy consumption in the nano-network using RL. This gap in the literature motivates our investigation to propose an adaptive channel bandwidth routing algorithm in nano-networks using RL to overcome energy consumption issues.

### C. *Contributions and Problem Statement*

The increasing prevalence of nano-networks, composed of nano-nodes with limited processing power, storage capacity, and energy resources, necessitates the development of low-complexity machine-learning algorithms. These algorithms must be able to operate efficiently within the constraints of these nano-nodes. Our primary goal is to improve electromagnetic nano-networks' energy efficiency and signal quality. We achieve this through a comprehensive approach that optimises routing traffic and dynamically adapts channel bandwidth. This strategy minimises energy consumption in nano-nodes and reduces energy per pulse while maintaining reliable communication. We leverage reinforcement learning (RL) to determine the optimal routing path (single or multi-hop) and sub-channel bandwidth allocation, ultimately minimising the total network energy consumption. Each nano-node will have an offline trained Q-table with state and action pairs, and it could select the optimal action for its state. Suppose the optimal action leads to any invalid state for any reason, such as the nano-node running out of energy or unavailable channel bandwidth. In that case, the nano-node does not need to re-run the learning algorithm again, where this recured more processing and energy resources. We developed a novel algorithm to enable the offline learned nano-node to skip this action and select the second optimal action from the available actions in the Q-table instead of re-running the algorithm again.

The primary achievements of this research are as follows:

- We developed an RL model using a Q-learning algorithm for energy-efficient THz nano-networking based on adaptive channel bandwidth allocation.
- We also optimise traffic routing in nano-networks, considering the entire path from nano-sensors to nano-routers.

### D. *Organisation*

The remainder of this paper is structured as follows: Section II overviews the Reinforcement Learning (RL) model, including its architecture and mathematical formulation. Section III details the proposed model, encompassing the environment and implemented policies. Section IV presents the results obtained from training and testing the model. Finally, Section V concludes the paper.

## II. REINFORCEMENT LEARNING (RL) ALGORITHM

Machine learning (ML), a branch of computer science, tackles practical problems by constructing statistical models from datasets. These models enable machines to learn and solve problems without explicit programming. Learning algorithms are applied to training sets, which consist of input samples. ML paradigms include supervised, semi-supervised, unsupervised, and reinforcement learning [20], [21]. Reinforcement learning (RL), a subfield of machine learning, involves an agent interacting with an environment, as shown in Fig. 1. Reinforcement learning (RL) stands out as a particularly exciting area within ML due to its ability to train agents through trial and error in dynamic environments [22] The agent perceives its state through feature vectors and can execute actions that lead to different rewards and state transitions. RL focuses on learning the optimal actions or sequences of actions that maximise cumulative rewards [23]. This feedback loop allows the agent to learn which actions are most likely to lead to positive outcomes. The objective of an RL algorithm is to discover a policy that maps state features to optimal actions, where optimality is defined as maximising the expected average reward.

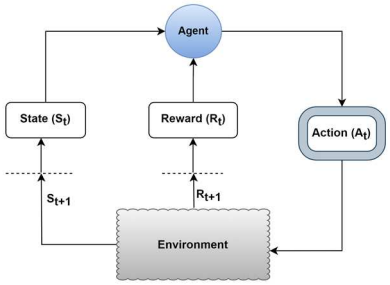


Fig. 1. Conceptual illustration of the reinforcement learning process.

Q-learning, a model-free reinforcement learning (RL) algorithm, facilitates the learning and optimisation of an agent's behaviour through iterative interactions with its environment [24]. Each iteration of the learning process is termed an (episode), where increasing the number of episodes enhances the agent's ability to learn from diverse reward structures and decision-making contexts. In each episode, the agent observes its current state ($S_t \in S$), takes an action ($A_t \in A$), and receives a reward ($R(S_t, A_t)$) from the environment. This cycle repeats as the agent transitions to the next state ($S_{t+1}$). Following each episode, the accumulated rewards and penalties for each state-action pair are stored in the Q-table, a reference table used by the agent to guide future decisions [25]. The agent's objective is to learn a policy, which is a behavioural rule that maximises its cumulative reward. The definition of the Q-learning algorithm can be represented as [24]

$$\mathrm{Q}_{t+1}(S_t, a_t) \leftarrow \mathrm{Q}_t(S_t, a_t) + \alpha \left[ R_{t+1} - \mathrm{Q}(S_t, a_t) + \gamma \max_a \mathrm{Q}_t(S_{t+1}, a_t) \right] \quad (1)$$

where $S_t$ represents the current state of the environment, $a_t$ is the action taken by the system, α is the learning rate $0 \leq \alpha \leq 1$, γ is the discount factor $0 \leq \gamma \leq 1$, $R_{t+1}$ is the received reward, $t$ in the number of episodes, $Q(S_t, A_t)$ is the Q value, which is dynamically adjusted based on the agent's initial interactions with the environment and $\max_a Q_t(S_{t+1}, a_t)$ is the highest expected future reward. The learning rate (α) controls the balance between prioritising past knowledge (low α) and emphasising new information (high α). However, the discount factor (γ) in Q-learning determines the importance of future rewards, with higher values emphasising long-term rewards and lower values prioritising immediate rewards [18].

## III. System Model

### A. Environment

This section presents the Q-learning algorithm based on reinforcement learning (RL) for optimising resource allocation in nano-networks. Our model assumes that a single nano-sensor transmits its data to the nano-router. To minimise energy consumption, the nano-sensor must select the optimal data transmission sub-channel bandwidth, considering both single-hop and multi-hop communication strategies while adhering to capacity and energy constraints. A customised OpenAI Gym environment was constructed to enable the training and evaluation of reinforcement learning algorithms on a nano-node, utilising the parameters outlined in Table I. The proposed model is based on the following assumptions:

- A network is set up with $|N|$ nano-nodes distributed across an area $(D_{max} \times D_{max})$ meter. Within the network of $|N|$ nano-nodes, one is dedicated to sensing nano-node that generates the traffic demand based on some collected data, another acts as a routing nano-node as the destination of the traffic demand, while the remaining nano-nodes are assumed to be relay nano-nodes $|\rho|$ for extending the network's reach $(|N| - 2)$. Our model utilises a centralised topology [3], where all nano-sensors within a cluster transmit their data to a central nano-router.
- Our model incorporates $L$ sub-channel options, each with a different bandwidth, ranging from 6.25 GHz to 100 GHz. Both the transmitter and receiver utilise arrays of graphene-based plasmonic nano-antennas [26], with each antenna capable of tuning to a specific central frequency and operating within a designated sub-channel bandwidth.

Q-learning, a reinforcement learning technique, can identify the optimal policy for selecting actions and maximising rewards [23]. As explained in Fig. 1, three main variables are incorporated into the Q-learning algorithm based on feedback received: (1) Discrete agent action (a) indicating the next hopping node; (2) Discrete environment state (S), which indicates the packet's position after each action; (3) Discrete environmental reward (R), which is a cumulative reward calculated during each action. In our model, the agent is the data packet that must be sent to the destination. The nano-sensor could select one of the $(|N| - 2) \times L$ possible actions to send to the destination because it could send to any of the $|N| - 2$ nodes in the cluster using one of $L$ available sub-channel bandwidths. The state is the packet position during its journey from the nano-sensor to the nano-router. This process can be represented by a sequence of states, actions, and rewards: $[S_0, a_0, R_1, S_1, a_1, R_2, S_2, a_2, R_3, ...]$. This sequence shows how the agent transitions between states and receives rewards for its actions.

TABLE I REINFORCEMENT LEARNING MODEL ENVIRONMENT PARAMETERS

| Symbol | Value and Unit | Description |
|---|---|---|
| $PT$ | 1 µW [27], [28], [29] | Nano-nodes transmit power |
| $PP$ | 140 nW [30] | Nano-processor processing power consumption |
| $PS$ | 50 nW [30] | Nano-sensor sensing power consumption |
| $TS$ | 2 µsec [27], [30] | The time between consecutive pulses |
| $EC_{max}$ | 800 pJ [31] | Energy storage capacity of the nano-capacitor |
| $D_{max}$ | From 1 mm to 90 mm | The maximum cluster transmitting distance. |
| F | 150 GHz | Central frequency |
| $BW$ | 100GHz, 50GHz, 25GHz, 12,5GHz, 6.25GHz | Set of the bandwidth options available for the link between any pair of nano-nodes |
| $Pb$ | 1 (worst-case scenario) | The probability of sending ones |
| $BZ$ | $1.38 \times 10^{-23}$ J/K | Boltzmann constant |
| $T_o$ | 296 K | Reference temperature of the medium |
| $C_o$ | $2.9979 \times 10^8$ m/s | Speed of light |
| $DM_{s,d}$ | 40 bits (5 bytes) [30] | The demand between the nano-sensor and the nano-router |
| $\varphi$ | $0 \leq \varphi \leq 1 \times 10^{-3}$ | Nano-nodes processing/sensing energy weighting parameter |
| $\alpha$ | 0.1 | Learning Rate |
| $\gamma$ | 0.9 | Discount Rate |
| $\varepsilon$ | 0.1 | Greedy rate |

### B. NETWORK INDEX

***Definition 1 (Distance Between Nano-Nodes):*** the distance between any two nano-nodes is calculated using the formula below:

$$d_{t,t+1} = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \quad (2)$$

Where $(x_{t+1}, y_{t+1})$ is the coordinate of the next node (next state), while $(x_t, y_t)$ is the coordinate of the current node (current state).

***Definition 2 (path-loss Between Nano-Nodes):*** the path-loss between any two nano-nodes is calculated using the formula below:

$$PL_{t,t+1} = \left(\frac{4\pi \times F \times d_{t,t+1}}{C_0}\right)^2 \quad (3)$$

***Definition 3 (Noise Between nano-Nodes):*** total noise power spectral density between any two nano-nodes, which includes the molecular absorption noise [1], [32] and system noise as an additional thermal like factor [28].

***Definition 4 (Channel_capacity between nano-Nodes):*** maximum channel capacity for the link between current nano-node ($N_i$) and the next node ($N_{i+1}$) at the bandwidth option $bw \in \mathrm{B}W$, where

$$C_{t,t+1} = bw_t \times \log_2\left(1 + \frac{\mathrm{PT}}{PL_{t,t+1} \times bw_t \times N\,(\mathrm{d})}\right) \quad (4)$$

$N\,(d)$ is the total noise power spectral density. In addition, Eq. (5) guarantees that the traffic in each link will not exceed the link channel rate.

$$\frac{1}{TS} \leq C_{t,t+1} \quad (5)$$

***Definition 5 (Energy-Based Path):*** the total energy consumption at each nano-node consists of two parts, as explained in the equations below,

$$ET_t = \left[\left(\frac{PT}{bw_t} \times \mathrm{Pb}\right) + (\varphi \times TS \times (PP + PS))\right] \quad (6)$$

$$ER_t = \left[\left(\frac{PT}{bw_t} \times 0.1\right) + (\varphi \times TS \times (PP + PS))\right] \quad (7)$$

where $ET_t$ the energy consumption for the transmitting and relay nano-node, while $E\mathrm{R}_t$ the energy consumption for the receiving and relay nano-node. In addition, Eq. (8) guarantees that the energy consumed by a nano-node remains within the limits of its nano-capacitor's maximum storage capacity.

$$(ET_i + ER_i) \leq \mathrm{EC}_{max} \quad (8)$$

Table II outlines the rewards and penalties employed by our proposed RL model during the training and testing phases of nano-node optimisation.

TABLE II. REINFORCEMENT LEARNING MODEL REWARDS AND PENALTIES DEFINITION

| Rewards or Penalty | Description |
|---|---|
| Destination Reward ($\boldsymbol{\beta}_r^D$) | The agent is awarded when it reaches the destination. |
| Energy Reward ($\boldsymbol{\beta}_r^{DE}$) | The agent is awarded when it consumes less energy. |
| The previous node visited penalty ($\boldsymbol{\beta}_p^V$) | The agent incurs a penalty upon revisiting a previously visited node. |
| Not Delivered Penalty ($\boldsymbol{\beta}_p^{ND}$) | The agent is penalised for not reaching the destination. |
| Exceed the maximum channel capacity ($\boldsymbol{\beta}_p^C$) | Transmission through a link exceeding the maximum channel capacity results in a penalty for the agent. |
| Exceed the maximum energy capacity ($\boldsymbol{\beta}_p^E$) | A penalty is incurred by the agent when its energy consumption surpasses the maximum storage capacity of the nano-capacitor. |

**Algorithm 1:** Q-Learning Minimum Energy Consumption (**QL-MEC**) Algorithm

**Input:** Nano-Nodes information ( initial coordinate)
**Output:** Q-table

**%Initialize network;**
Learning rate $\boldsymbol{\alpha} \in (\mathbf{0}, \mathbf{1})$, Discount factor $\boldsymbol{\gamma} \in (\mathbf{0}, \mathbf{1})$, Greedy rate $\boldsymbol{\varepsilon} \in (\mathbf{0}, \mathbf{1})$
No. of Episodes
Zeros-matrix → Q-table
**% Run the routing algorithm;**
**while** Episode < No. of Episodes **do**
**Initialise:** Destination coordinate (Node 30)
  Source coordinate (Node1),
  Reset Current State ($\boldsymbol{S_t}$)
  Visited_state matrix → zero
  Maxinum steps =50 (accepted trial steps )
  Reset Total Rewards to 0
  Reset Total Energy to 0
  **% choose action**
  **If** generated random variable≤ Greedy rate **then**
    Generate random next action
  **else**
    Consider this Q-table value for the next action
  **end if**
  **while** step < Maximum steps **do**
    Generate the next step ($\boldsymbol{S_{t+1}}$) from the action
    Select one bandwidth depending on the action
    **if** $\boldsymbol{S_{t+1}}$ has been selected **or** equal to $\boldsymbol{S_t}$ **then**
      the agent gets a penalty = $\boldsymbol{\beta_p^V}$
    **end if**
    Calculate the distance between $\boldsymbol{S_t}$ and $\boldsymbol{S_{t+1}}$
    Calculate the channel capacity for the link $\boldsymbol{S_t}$,$\boldsymbol{S_{t+1}}$
    Calculate the total energy consumption
    **if** (7) is not satisfied **then**
      the agent gets a penalty = $\boldsymbol{\beta_p^C}$
    **end if**
    **if** (10) is not satisfied **then**
      the agent gets a penalty = $\boldsymbol{\beta_p^E}$
    **else**
      the agent gets a reward = $\boldsymbol{\beta_r^{DE}}$
    **end if**
    **If** the agent reached the destination **then**
      the agent gets a reward = $\boldsymbol{\beta_r^D}$
    **else**
      the agent gets a penalty = $\boldsymbol{\beta_p^{ND}}$
    **end if**
    update the Visited_state matrix, $\boldsymbol{S_t}$ and $\boldsymbol{S_{t+1}}$
  **end while**
  % Update Q-table with states, rewards, selected bandwidth, and total energy
**end while**

Our RL algorithm, named Q-Learning Minimum Energy Consumption (QL-MEC), is described in detail in Algorithm 1. The agent is trained to find the optimal path to the destination by receiving rewards for correct decisions and penalties for incorrect ones. The optimal path means the optimal next step with optimal sub-channel bandwidth allocation, which means the lowest energy consumption. This policy encourages exploration during training, allowing the agent to learn from its mistakes. Reaching the destination with the lowest energy consumption earns the highest reward while consuming more energy results in a penalty. During training

episodes, the algorithm also monitors channel capacity and the maximum energy remaining in the nano-node. After each step, the agent verifies if the next state's connectivity surpasses the channel capacity constraint (5) and energy storage constraint (8). If one or both are not satisfied, a penalty $\boldsymbol{\beta}_p^C$ or $\boldsymbol{\beta}_p^E$ or both are incurred. Furthermore, to prevent infinite routing loops, the agent incurs a penalty $\boldsymbol{\beta}_p^V$ for revisiting a previously visited node within the same episode. Conversely, reaching the destination nano-router yields a substantial reward $\boldsymbol{\beta}_r^D$. As training progresses and Q-values are updated, the greedy rate ($\boldsymbol{\varepsilon}$) gradually decreases. This parameter balances exploration (random action selection) and exploitation (using learned Q-values) to guide the agent's decision-making for subsequent actions.

## IV. TRAINING AND EVALUATION RESULTS

In this paper, we considered one nano-node cluster with varying areas, ranging from 1 $mm^2$ to 70 $mm^2$. Each cluster consists of one nano-sensor, one nano-router, and 28 nano-relays. The analysis considers a bandwidth range of 0.1 to 1 THz [27], [30]. Five different bandwidth options are evaluated, as shown in TABLE I. Our model assumes the symbol probability of logical "1" is 1, representing the worst-case scenario in terms of energy consumption. By adjusting the energy weighting parameter ($\varphi$), we can analyse how the proportion of energy consumed by processing/sensing operations affects the optimal selection strategy. When $\varphi$ =0, processing/sensing energy is considered negligible compared to transmission/reception energy. Conversely, when $\varphi > 0$, processing/sensing energy becomes the dominant factor.

Firstly, we evaluated the QL-MEC model using the assumptions and parameters explained beforehand. Figure. 2 shows that heavy penalties for the agent characterise initial training episodes, as the Q-model is still unfamiliar with the environment. This highlights the need for further training, given the various possible states and actions. During each state, the agent updates its current state, and next state, total energy consumption to move to the next step, selects channel bandwidth, receives rewards or penalties based on its actions, and refines the Q-model accordingly. The agent learns to navigate the environment effectively and efficiently through repeated episodes and by utilising the knowledge stored in the Q-model. The total reward increases as the number of episodes increases, indicating that the agent is learning to perform the task more effectively. By the end of the training period, the agent had learned to perform the task quite well and received a high total reward. This suggests that the agent has learned a good policy for choosing actions in the environment.

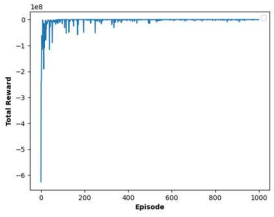


Fig. 2. Total Rewards vs Episodes by QL-MEC Algorithm.

We test our QL-MEC model to evaluate the performance of our trained agents. The trained agent improve their decision-making through iterative training, leading to increased rewards and reduced penalties. The effectiveness of the trained nano-nodes is assessed using average rewards and penalties gathered during training. The trained nano-nodes were tested in environments with comparable dimensions to assess their performance. Figure. 3 shows training the model when the maximum cluster size is about 20 mm, using the same simulation setting discussed in section IV. Figure. 3(a) and Figure. 3(b) show that the nano-sensor (S1) trained from its environment where its reward has been increased while the total network energy consumption is decreased. At the end of 300000 episodes, the total reward is 998 while the total energy consumption is $1.76 \times 10^{-15}$ joule. It was noticed from the training cycle that S1 selects multi-hops for the first 2000 episodes, then selects the single-hop, albeit with the lowest channel bandwidth of up to 10000 episodes. The model then tries to minimise energy consumption by using higher channel bandwidth. When the number of episodes hits 100000, the QL-MEC model selects the optimal channel bandwidth (25 GHz) option for the link from S1 to the nano-router to minimise the total network energy consumption. This optimal selection is saved in the Q-table and should be selected when the test cycle starts. We examined our model by running the test 100 times to check its accuracy. Figures 3(c)-(d) show that all 100 test samples select the optimal route with the optimal channel bandwidth selection, which means a higher reward and more energy-efficient selection.

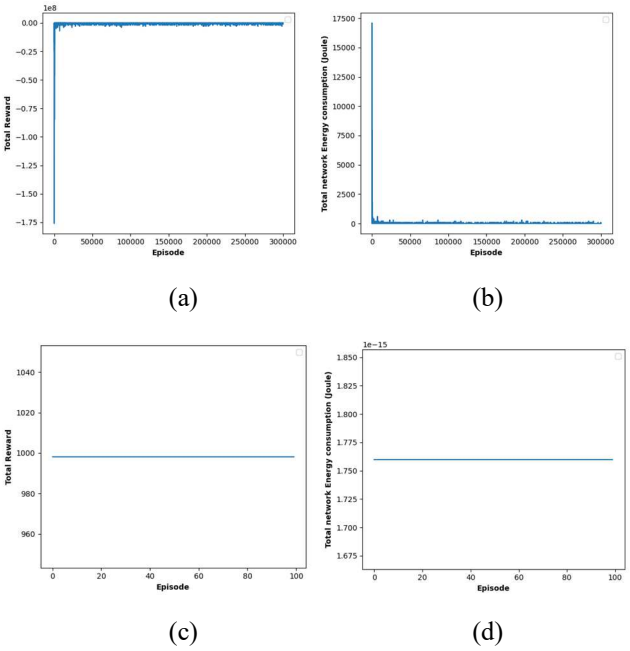


Fig. 3. Testing the QL-MEC model 100 times at the maximum cluster size of 20 mm. (a) Total reward during the 300000 training cycle. (b) Total energy consumption during the 300000 training cycle. (c) Total reward during the 100-test cycle. (d) Total energy consumption during the 100-test cycle.

Then, we test the QL-MEC Algorithm at different values of $D_{max}$ to show how that affects the optimal selection of sub-channel bandwidths and routing strategies. Figure. 4 shows

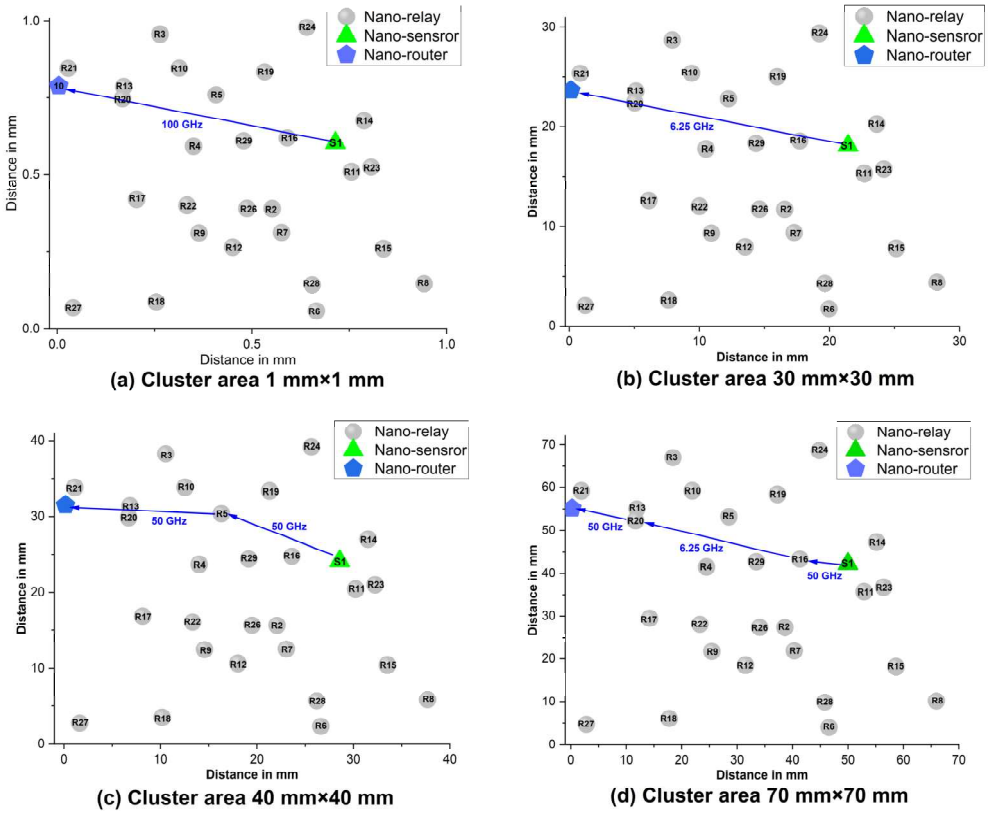


Fig. 4. Analysing network traffic flow under the QL-MEC model for short and long transmission distances with optimal channel bandwidth selections, where φ=0 and PT=1μW.

the optimal routing traffic and optimal channel bandwidth selection strategy for a nano-node cluster with a nano-sensor (S1), a nano-router, and several nano-relays at $\varphi = 0$ and at different $D_{max}$ values. At very short distances ($D_{max} = 1\ mm$), the S1 uses a single hop with the highest channel bandwidth (100 GHz) to send its traffic directly to the nano-router. This selection helps reduce the energy consumption for the pulse duration, which leads to decreased total network energy consumption. As the $D_{max}$ increases, the S1 tries to keep using a single hop by reducing the channel bandwidth to satisfy the channel capacity constraint (5). For example, the lowest channel bandwidth (6.25 GHz) is used by S1 to send its traffic directly to the nano-router when $D_{max}$ increased up to 30 $mm$, Fig. 4(b). Generally, the S1 uses a single hop with a lower channel bandwidth to achieve the required data rate. However, decreasing the channel bandwidth at a certain transmitting distance will not be an efficient solution to save energy, so multi-hop with higher channel bandwidth will be the best option. For example, S1 uses R5 to relay its traffic when $D_{max}$ is up to 40 $mm$, allowing S1 to reuse a higher channel bandwidth (50 GHz) to reduce energy consumption, as shown in Fig. 4(c). Increasing $D_{max}$ more than 40 $mm$, lower channel bandwidth will be used to satisfy the channel capacity constraint. Similarly, when $D_{max}$ is 70 $mm$, S1 uses R16 and R17 to relay its traffic, allowing it to reuse a higher channel bandwidth (50 GHz), Fig. 4(d). Overall, the optimal channel bandwidth selection strategy is a complex decision that depends on some factors, including the distance between nano-nodes and the availability of relay nodes. When the distance is short, a higher channel bandwidth can be used. However, a lower channel bandwidth may be needed to maintain a good signal-to-noise ratio when the distance is long. Additionally, using a relay node can allow for a higher channel bandwidth, which can save energy. Considering all these factors, the optimal channel bandwidth selection strategy can help ensure reliable communication while minimising energy consumption.

This section investigates the impact of incorporating processing/sensing units ($\varphi = 1 \times 10^{-3}$) on the overall energy consumption within the QL-MEC model. Figure. 5 illustrates a notable increase in total network energy consumption for the OL-MEC model equipped with processing/sensing units compared to the model without them. This rise is attributed to the inherent energy demands of these additional units. Despite the increased energy consumption, the trend of energy savings concerning the maximum cluster size remains consistent across both models. Specifically, the total network energy consumption for the OL-MEC model with $\varphi > 0$ exhibits a slight increase as the maximum cluster size expands up to 40 $mm$. Beyond this point, a significant surge in energy consumption occurs due to activating a relay node for traffic routing, which introduces additional processing/sensing energy requirements.

Furthermore, the analysis reveals that the OL-MEC model with $\varphi > 0$ utilises only one nano-relay with a lower channel bandwidth compared to the model with $\varphi = 0$, which employs two nano-relays when the maximum cluster size

reaches 70 mm. This observation underscores the energy inefficiency of multi-hop communication due to the substantial energy consumption associated with processing units within nano-relays. Our current work, while offering valuable insights, operates within certain constraints. The Q-table should be shared between all nano-nodes. A limited number of nano-nodes are assumed in our model, with just one nano-sensor. It is assumed that at multi-hops, different nano-nodes could use the same central frequency, and that could add an additional source of noise, which is not considered in this work.

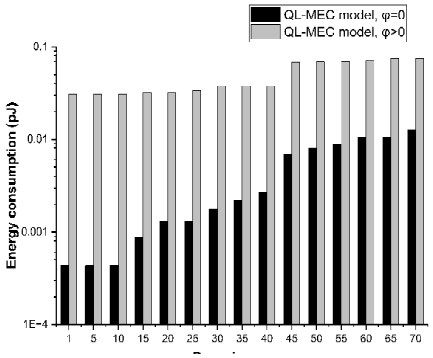


Fig. 5. The effect of adding the processing/sensing units ($\varphi = 1 \times 10^{-3}$) on the QL-MEC model when PT=1μW.

We've designed a lightweight, adaptive algorithm capable of running on individual nano-nodes to determine the optimal route with the best possible channel bandwidth. This algorithm operates in two stages. First, Algorithm 1 runs offline to generate a Q-table, which is then distributed to all nano-nodes. These nano-nodes utilize the pre-trained Q-table to select the most efficient actions for their tasks. However, resource limitations may occasionally render these initially selected actions infeasible. We avoid re-training the entire algorithm in such situations to conserve precious nano-node resources. Instead, as outlined in Algorithm 2, the nano-node temporarily stores the Q-value of the optimal action-state pair and sets it to zero within the table. This allows the selection of the following best action from the existing Q-table. Notably, the original Q-value for the previously infeasible action is restored after transitioning to the next state. This ensures its availability as a potential optimal action in the future, preventing the need for costly re-training of the algorithm. Figure. 6 demonstrates a scenario where node S1 chooses R29 as its routing path instead of R5, which was initially identified as the optimal relay at a distance $D_{max} = 40\ mm$, as depicted in Fig. 4(c). Interestingly, S1 utilizes a 100 GHz channel bandwidth for its connection to R29, compared to the 50 GHz bandwidth allocated for the now unavailable link to R5. This higher bandwidth allows for shorter pulse durations, contributing to energy savings. However, this adaptation comes with a trade-off. The channel bandwidth between R29 and the destination nano-node is reduced to 12.5 GHz, significantly lower than the 50 GHz bandwidth of the previously optimal R5 connection. Consequently, the overall network energy consumption experiences a slight increase when utilizing R29 instead of R5. Despite this slight energy increase, our model offers a valuable advantage: it conserves nano-node resources by allowing them to seamlessly select the second-best action without requiring computationally expensive re-training of the entire algorithm. This adaptability ensures efficient operation even when the initial optimal path becomes unavailable.

**Algorithm 2: Adaptive QL-MEC algorithm**

**Input:** Q-table
**Output:** optimal action
1 **%Initialize network;**
2 **while** the nano-sensor packet has not reached the destination, **do**
3 **% choose action**
4 Choose the best action from Q-table for the current state
5 **while** the selected action is not valid, **do**
6 Store the action in a temporary variable
7 Update the Q value of this action, state pair with zero
8 Choose the best action from the Q-table for the current state
9 **end while**
10 Go to the next state
11 Restore the original value of the invalid action/s
12 **end while**

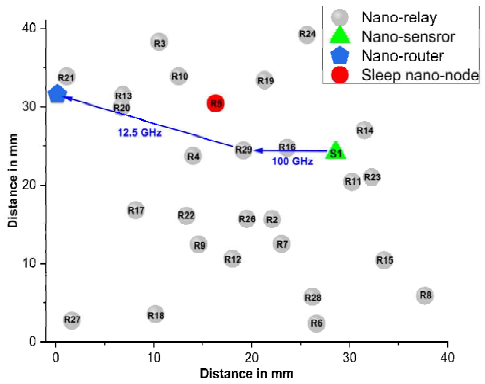


Fig. 6. Analysing network traffic flow under the adaptive QL-MEC algorithm with optimal channel bandwidth selections when nano-relay (R5) runs out of energy, where φ=0 and PT=1μW.

## V. CONCLUSION

We developed an energy-efficient Q-learning algorithm based on reinforcement learning (QL-MEC) and tested it at different cluster sizes. The algorithm is simulated using OpenAI Gym environment, where a nano-sensor is asked to send its traffic to the nano-router with minimum energy consumption. Our results show the model's robust performance in selecting the optimal traffic route, optimal channel bandwidth, and minimum energy consumption. The QL-MEC model uses single-hop with higher channel bandwidth at short transmitting distances to save energy. However, the higher channel bandwidth is selected as the transmitting distances increase until the single-hop transmission becomes invalid. At this point, the QL-MEC model tries to route the traffic through the relay node and reuse the higher channel bandwidth option selection to save energy. While the inclusion of processing and sensing units' energy consumption may impose limitations on the practicality of multi-hop communication across certain distances, our proposed model nonetheless exhibits characteristics of an

energy-efficient routing scheme. The model's adaptability extends to scenarios such as multi-nano-sensors sending simultaneously and optimising the sub-channel bandwidth allocation, which will be explored in future research endeavours.

## REFERENCES


[1] J. M. Jornet and I. F. Akyildiz, "Channel Modeling and Capacity Analysis for Electromagnetic Wireless Nanonetworks in the Terahertz Band," *IEEE Transactions on Wireless Communications,* vol. 10, pp. 3211-3221, 2011.

[2] I. T. Javed and I. H. Naqvi, "Frequency band selection and channel modeling for WNSN applications using simplenano," in *2013 IEEE International Conference on Communications (ICC)*, 9-13 June 2013 2013, pp. 5732

[3] M. Pierobon, J. M. Jornet, N. Akkari, S. Almasri, and I. F. Akyildiz, "A routing framework for energy harvesting wireless nanosensor networks in the Terahertz Band," *Wireless Networks,* vol. 20, no. 5, pp. 1169-1183, 2014/07/01

[4] J. Xu, R. Zhang, and Z. Wang, "An energy efficient multi-hop routing protocol for terahertz wireless nanosensor networks," in *International Conference on Wireless Algorithms, Systems, and Applications*, 2016: Springer,

[5] I. F. Akyildiz and J. M. Jornet, "Electromagnetic wireless nanosensor networks," *Nano Communication Networks,* vol. 1, no. 1, pp. 3-19, 2010/03/01/ 2010.

[6] I. F. Akyildiz, F. Brunetti, and C. Blázquez, "Nanonetworks: A new communication paradigm," *Computer Networks,* vol. 52, no. 12, pp. 2260-2279, 2008/08/22/ 2008.

[7] S. Mohrehkesh and M. C. Weigle, "Optimizing Energy Consumption in Terahertz Band Nanonetworks," *IEEE Journal on Selected Areas in Communications,* vol. 32, no. 12, pp. 2432-2441, 2014.

[8] S. Mohrehkesh and M. C. Weigle, "Optimizing communication energy consumption in perpetual wireless nanosensor networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, 9-13 Dec. 2013 2013, pp. 545-550.

[9] A. O. Balghusoon and S. Mahfoudh, "Routing Protocols for Wireless Nanosensor Networks and Internet of Nano Things: A Comprehensive Survey," *IEEE Access,* pp. 1-1, 2020.

[10] A. Oukhatar, M. Bakhouya, D. E. Ouadghiri, and K. Zine-Dine, "Probabilistic-Based Broadcasting for EM-based Wireless Nanosensor Networks," presented at the 15th International Conference on Advances in Mobile Computing & Multimedia, Salzburg, Austria, 2017.

[11] A. Tsioliaridou, C. Liaskos, S. Ioannidis, and A. Pitsillides, "CORONA: A Coordinate and Routing system for Nanonetworks," presented at the Proceedings of the Second Annual International Conference on Nanoscale Computing and Communication, Boston, MA, USA, 2015.

[12] C. C. Wang, X. W. Yao, W. L. Wang, and J. M. Jornet, "Multi-Hop Deflection Routing Algorithm Based on Reinforcement Learning for Energy-Harvesting Nanonetworks," *IEEE Transactions on Mobile Computing,* vol. 21, no. 1, pp. 211-225, 2022.

[13] C. Liaskos, A. Tsioliaridou, S. Ioannidis, N. Kantartzis, and A. Pitsillides, "A deployable routing system for nanonetworks," in *2016 IEEE International Conference on Communications (ICC)*, 22-27 May 2016 2016, pp. 1-6.

[14] A. Tsioliaridou, C. Liaskos, E. Dedu, and S. Ioannidis, "Stateless Linear-path Routing for 3D Nanonetworks," presented at the Proceedings of the 3rd ACM International Conference on Nanoscale Computing and Communication, New York, NY, USA, 2016.

[15] X.-W. Yao, Y.-C.-G. Wu, and W. Huang, "Routing techniques in wireless nanonetworks: A survey," *Nano Communication Networks,* vol. 21, p. 10025, 2019.

[16] F. Afsana, M. Asif-Ur-Rahman, M. R. Ahmed, M. Mahmud, and M. S. Kaiser, "An Energy Conserving Routing Scheme for Wireless Body Sensor Nanonetwork Communication," *IEEE Access,* vol. 6, pp. 9186-9200, 2018.

[17] H. Yu, B. Ng, and W. K. G. Seah, "TTL-Based Efficient Forwarding for Nanonetworks With Multiple Coordinated IoT Gateways," *IEEE Internet of Things Journal,* vol. 5, no. 3, pp. 1807-1815, 2018.

[18] H. Tan, T. Ye, S. u. Rehman, O. ur Rehman, S. Tu, and J. Ahmad, "A novel routing optimization strategy based on reinforcement learning in perception layer networks," *Computer Networks,* vol. 237, p. 110105, 2023/12/01/ 2023.

[19] A. J. Garcia-Sanchez, R. Asorey-Cacheda, J. Garcia-Haro, and J. L. Gomez-Tornero, "Dynamic Multihop Routing in Terahertz Flow-Guided Nanosensor Networks: A Reinforcement Learning Approach," *IEEE Sensors Journal,* vol. 23, no. 4, pp. 3408-3422, 2023.

[20] R. Boutaba *et al.*, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," *Journal of Internet Services and Applications,* vol. 9, no. 1, p. 16, 2018/06/21 2018.

[21] A. Burkov, *The hundred-page machine learning book.* Andriy Burkov Quebec City, QC, Canada, 2019.

[22] M. M. Qazzaz, S. A. Zaidi, D. McLernon, A. Salama, and A. A. Al-Hameed, "Low complexity online RL enabled UAV trajectory planning considering connectivity and obstacle avoidance constraints," in *2023 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2023: IEEE, pp. 82-89.

[23] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," presented at the Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 2016.

[24] C. J. C. H. Watkins and P. Dayan, "Technical Note: Q-Learning," *Machine Learning,* vol. 8, no. 3, pp. 279, 1992.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] L. Zakrajsek, D. Pados, and J. Jornet, *Design and performance analysis of ultra-massive multi-carrier multiple input multiple output communications in the terahertz band* (SPIE Commercial + Scientific Sensing and Imaging), 2017.

[27] S. Canovas-Carrasco, A.-J. Garcia-Sanchez, F. Garcia-Sanchez, and J. Garcia-Haro, "Conceptual design of a nano-networking device," *Sensors,* vol. 16, no. 12, p. 2104, 2016.

[28] I. Llatser *et al.*, "Scalability of the Channel Capacity in Graphene-Enabled Wireless Communications to the Nanoscale," *IEEE Transactions on Communications,* vol. 63, no. 1, pp. 324-333, 2015.

[29] P. M. Shree, T. Panigrahi, and M. Hassan, "Classifying the Order of Higher Derivative Gaussian Pulses in Terahertz Wireless Communications," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 9-13 Dec. 2018 2018, pp. 1-6.

[30] S. Canovas-Carrasco, A.-J. Garcia-Sanchez, and J. Garcia-Haro, "A nanoscale communication network scheme and energy model for a human hand scenario," *Nano Communication Networks,* vol. 15, pp. 17-27, 2018.

[31] J. M. Jornet, "A joint energy harvesting and consumption model for self-powered nano-devices in nanonetworks," in *2012 IEEE International Conference on Communications (ICC)*, 10-15 June 2012 2012, pp. 6151-6156.

[32] S. Paine, *The am atmospheric model*. 2019.