REAM: A Reinforcement Learning-based Energy-Efficient and Adaptive Multi-Modal Routing Protocol for Underwater Acoustic Networks

Rabia Qadar[1], Waleed Bin Qaim[2], Bo Tan[1], Jari Nurmi[1]
Email: {rabia.qadar, waleed.binqaim, bo.tan, jari.nurmi}@tuni.fi
[1]Tampere Wireless Research Centre, Unit of Electrical Engineering,
[2]Unit of Computing Sciences,
Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland.

***Abstract*—Improving reliability in dense networks with heavy data traffic often increases energy consumption in contentious aquatic networks. This heightened energy utilization primarily originates from escalated activities, including constant listening for incoming packets and retransmissions initiated due to packet loss. To address the challenging energy demand in Underwater Acoustic Sensor Networks (UASNs), we propose a Reinforcement learning (RL)-based Energy-efficient and Adaptive Multi-modal routing protocol abbreviated as REAM, that integrates Q-learning with multi-modal communication to enhance energy efficiency in underwater networks by adaptively selecting the optimal mode for packet transmission. We compare the performance of two variants of the proposed REAM protocol—REAM-MM, which uses two modems, and REAM-SM, which uses a single modem, against two other state-of-the-art protocols namely QELAR and MARLIN-Q. Our results demonstrate the effectiveness of using multiple modems instead of a single modem in reducing energy consumption and improving reliability. We specifically show that REAM-MM reduces energy consumption per bit by up to 81.2%, 79.6%, and 72.9% under low, medium, and high traffic scenarios, respectively, compared to the best-performing alternative, MARLIN-Q. REAM-MM achieves a consistently comparable Packet Delivery Ratio (PDR) to MARLIN-Q and a higher PDR than QELAR and REAM-SM. Additionally, it maintains the lowest energy consumption under all traffic conditions and dense networks, where nodes utilize a combination of high-frequency and low-frequency acoustic modems.**

***Index Terms*—Underwater communication, multi-modal acoustics, routing protocol, reliability, energy efficiency, Q-learning.**

Foundation for Economic Education, Finland (Grant number: 240031) funded the work of Waleed Bin Qaim at Haltian. This work is collaborated with the DIOR project that has received funding from the European Union's MSCA RISE programme under grant agreement No. 10100828.

I. Introduction

Underwater Acoustic Sensor Networks (UASNs) play a crucial role in ocean explorations by enabling scientists, researchers, and divers to exchange information, coordinate activities, and collect oceanic data in vast and challenging underwater environments using acoustic communication. These underwater operations are often mission-critical and require reliable and continuous communication. Failure or disruption in communication due to limited battery capacity in underwater nodes can have severe consequences, including data loss, compromised operations, or even safety hazards [1]. Additionally, underwater devices, such as sensors and Autonomous Underwater Vehicles (AUVs), often operate in deep, remote, and inaccessible locations, making it difficult to recharge or replace their batteries frequently. The limited battery capacity restricts the amount of energy available for communication activities such as routing while the environmental factors impact the performance and reliability of acoustic communication [2].

In this work, we propose a Reinforcement learning (RL)-based Energy-efficient and Adaptive Multi-modal routing protocol (REAM) that addresses the limited battery capacity of underwater devices and effectively handles the network dynamics using multi-modal acoustic communication. REAM enhances energy efficiency by integrating a high-frequency acoustic modem for shorter distances between nodes and a dependable low-frequency acoustic modem is employed for longer node-to-node distance. This intelligent modality selection is based on specific network load (number of active nodes), topology (arrangement of nodes),

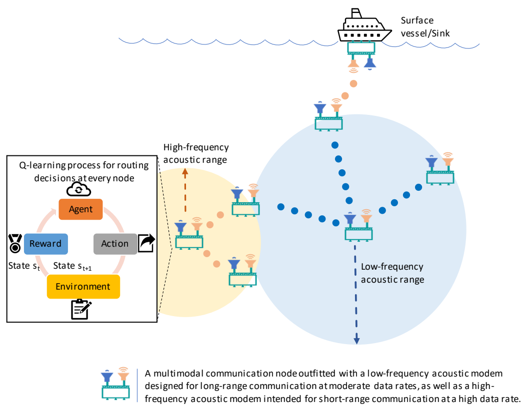


Fig. 1: Underwater multi-modal communication network

and available energy resources. It offers great flexibility in range, data rates, energy consumption, and adaptability to channel variations. Fig. 1 depicts the system scenario assumed in this work.

Our proposed method uses the model-based Q-Learning approach to determine the next potential relay and communication mode for routing a packet, at each node. It adapts physical layer parameters, such as communication range and modulation scheme, to reflect channel conditions. Furthermore, the protocol optimizes Medium Access Control (MAC) layer parameters, such as contention window (CW) size and time slot duration, to enhance routing. The CW size defines the range of time slots from which a node randomly selects its back-off time before transmitting a packet. A larger CW reduces collisions and retransmissions, lowering energy consumption but increasing wait times, while a smaller CW minimizes wait times but raises collision risk and energy costs. The optimal CW size depends on network density and load; dense networks require larger CWs to avoid collisions, whereas lighter loads benefit from smaller CWs. Our analysis focuses on identifying the optimal CW size under varying network conditions for maximum reliability and energy efficiency.

The rest of the paper is organized as follows. Section II describes the mathematical framework behind REAM. Section III explains the protocol mechanism and section IV details the simulation environment. Section V discusses the results and important findings. Finally, section VI concludes the paper.

II. The REAM Protocol

The proposed algorithm enables underwater devices to dynamically switch between communication modalities based on the node's residual energy, and transmission range. By selecting the most energy-efficient modality for a given routing task, REAM enables devices to optimize battery usage and extend the network's operational lifetime while maintaining reliability and interconnectivity.

A. The Markov Decision Process (MDP) Framework

We simulate a multi-modal UASN as a multi-agent system to optimize individual routing decisions through information exchange between immediate neighbor nodes. Each sensor node has two modems and is regarded as an ***agent***. The complete operation is modeled as a MDP [3]. As illustrated in Fig. 2, an MDP is defined by the quintuplet $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \pi)$, where $\mathcal{S}$ is the finite set of states, $\mathcal{A}$ is the finite set of actions, $\mathcal{P}$ is the state transition probability, $\mathcal{R}$ is the reward of the taken action, and π is the optimal policy. The definitions of the quintuple in this work are as follows:

1) State: The condition of an agent, i.e., a node n_i holding a packet is known as the ***state*** of the system. The finite set of states is mathematically defined as:

$$\mathcal{S} = \{s_1, s_2, ..., s_n\}, \tag{1}$$

where s_i represents each possible state in the system, and n represents the aggregate count of distinct states, i.e., the total number of nodes in the network. The state s_i changes when n_i forwards a packet to one of its one-hop neighbors n_{ij}.

2) Action: A node perceives the current network state and makes an informed routing decision called an ***action***. Let M be the set of modems containing a Low-Frequency modem (LF_m) and a High-Frequency modem (HF_m), and N be the set of neighbors of node n_i. Then a finite set of actions is defined as:

$$\mathcal{A}_i^{m_{ij}}(s) = \{a = \langle n_{ij}, m_{ij} \rangle \mid m_{ij} \in M, n_{ij} \in N\}, \tag{2}$$

where $a = \langle n_{ij}, m_{ij} \rangle$ is the joint action of forwarding a packet from node n_i to the next hop neighbor n_{ij} using modem m_{ij}. When the relay nodes forward the packet either directly to the sink or the next forwarder, the state changes and the node calculates the Q-value $Q^{\pi}(s_t, a_t)$ which defines the goodness of the action $a_t = \langle n_{ij}, m_{ij} \rangle$ in the state s_t under a policy π at time t given as:

$$Q^{\pi}(s_t, a_t) = r_t + \gamma \sum_{s_{t+1} \epsilon S} P_{s_t s_{t+1}}^{a_t} V^{\pi}(s_{t+1}), \tag{3}$$

where r_t is the immediate reward, $\gamma \in (0, 1]$ denotes the discount factor regulating the effect of the future rewards (term encapsulated within

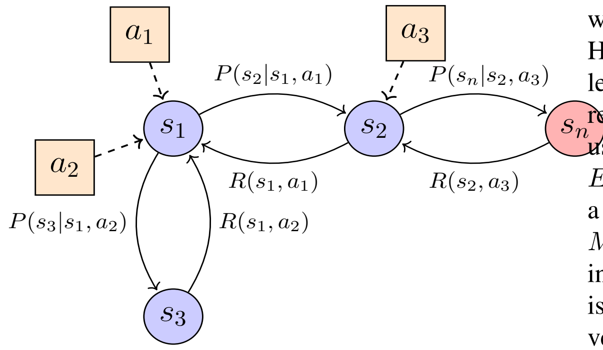


Fig. 2: Illustration of a Markov Decision Process (MDP) showing actions, state transitions, and the associated rewards.

$\sum_{s_{t+1}}$). Hence, the maximum Q-value attained is represented as $V^\pi(s)$ is given as:

$$V^\pi(s) = \max_a Q^\pi(s, a). \tag{4}$$

3) State Transition Probability: The state transition probability, $P_{s_i s_j}$, is expressed as the ratio of the number of successfully received packets δ to the total number of sent packets $\sum_{i=1}^n T_i$, symbolizing the likelihood of successful transmission between states s_i and s_j:

$$P^a_{s_i s_j} = \frac{\delta}{\sum_{i=1}^n T_i}. \tag{5}$$

However, if the packet fails to reach the next-hop neighbor, the state remains unchanged, i.e., s_i, and the respective transition probability is given as:

$$P^a_{s_i s_i} = 1 - P^a_{s_i s_j}. \tag{6}$$

4) Reward: The quantified outcome for an action a taken in state s is called the ***reward***. If the packet is successfully received, the obtained reward is defined as:

$$R = -r_o - \alpha \times \{\mathrm{E}(a)\}, \tag{7}$$

where r_o is the constant cost, $\alpha \in (0, 1)$ is the weight of the energy-related cost of the two nodes involved in the transmission and is set to 1. $\mathrm{E}(a)$ is the feedback part of the reward function R, known as the cost of energy. It quantifies the residual energy as a fraction of the initial energy after deducting energy expenditures for LF_m and HF_m given as:

$$\mathrm{E}(a) = \left(1 - \frac{E^{n_i}_{\text{res}} - \left(M_1 E^{LF_m}_s + M_2 E^{HF_m}_s\right)}{E^{n_i}_{\text{ini}}}\right) + \left(1 - \frac{E^{n_{ij}}_{\text{res}} - \left(M_1 E^{LF_m}_r + M_2 E^{HF_m}_r\right)}{E^{n_{ij}}_{\text{ini}}}\right), \tag{8}$$

where the initial energy E_{ini} is equal for all nodes. However, the residual energy E_{res} is the energy left after the transmission at time t. $E^{LF_m}_s$, $E^{HF_m}_s$ represent the energy utilized in sending a packet using LF_m and HF_m, respectively, while $E^{LF_m}_r$, $E^{HF_m}_r$ denote the energy consumption for receiving a packet through LF_m and HF_m, respectively. M_1, M_2 are constants indicating the modem used in the respective action. For instance, when LF_m is used, it sets $M_1 = 1$ and $M_2 = 0$, and vice versa when HF_m is used. Nodes with more residual energy E_{res} will have a lower cost and are more likely to be selected as potential forwarders.

In case of unsuccessful transmission, the added cost of energy due to k retransmissions is considered, i.e., $k \times E_s$, and the respective cost of energy in equation (8) for unsuccessful transmission is updated as:

$$\mathrm{E}(a) = \left(1 - \frac{E^{n_i}_{\text{res}} - (M_1 E^{LF_m}_s + M_2 E^{HF_m}_s) \times k}{E^{n_i}_{\text{ini}}}\right) + \left(1 - \frac{E^{n_{ij}}_{\text{res}}}{E^{n_{ij}}_{\text{ini}}}\right). \tag{9}$$

Since the next forwarder did not receive the packet, the term E_r is not added to the cost.

By putting equations (8) and (9) in (7), we derive the rewards $R^a_{s_i s_j}$, $R^a_{s_i s_i}$ for the successful and unsuccessful transmission, respectively. The total reward can be defined as:

$$r^{m_{ij}}_i = P^a_{s_i s_j} R^a_{s_i s_j} + P^a_{s_i s_i} R^a_{s_i s_i}. \tag{10}$$

According to equations (3) and (4), the action-utility function $Q^\pi(s_i, a)$ and optimal strategy $V^\pi(s)$ are given as:

$$Q^\pi(s_i, a) = r^{m_{ij}}_i + \gamma\left(P^a_{s_i s_j} V^\pi(s_j) + P^a_{s_i s_i} V^\pi(s_i)\right), \tag{11}$$

$$V^\pi(s) = max_a \left(Q^\pi(s_i, a)\right). \tag{12}$$

The Q-values are iteratively updated during the learning process to converge toward the optimal policy π that balances energy efficiency and reliability.

5) Optimal Policy: The optimal policy $\pi(s)$ denotes the strategy to maximize the state-action pair $Q^\pi(s_t, a_t)$ by selecting an action a in the state s that minimizes the cost of energy $\mathrm{E}(a)$ given as:

$$\pi(s) = \operatorname*{argmin}_a [\mathrm{E}(a)]. \tag{13}$$

III. Protocol Mechanism

The proposed REAM protocol consists of three key phases: neighbor and mode discovery, learning, and adaptation. The following subsections provide a detailed explanation of each phase in the algorithm.

A. Classification of Neighbors and Modes

In the first phase, all the nodes in the network obtain the 8-bit address, location coordinates x_i, y_i, z_i[1], mode of connection m_{ij}, initial Q-values Q_i, and residual energy $E_{res}^{n_{ij}}$ of their one-hop neighbors n_{ij} through the exchange of broadcast packets (Lines 4 to 7 in Algorithm 1). Nodes classify n_{ij} as high-frequency (HF) for short-range and low-frequency (LF) for long-range communication based on proximity.

Algorithm 1 Discover Neighbors and Modes

1: **Input:** x_i, y_i, z_i
2: **Output:** N, M
3: **Initialize:** $Q_i \leftarrow 0, E_{res}^{n_{ij}} \leftarrow E_{\text{ini}}$
4: **for all** Nodes **do**
5:     Broadcast p from n_i using m_{ij}
6:     $N \leftarrow \{n_{ij} \mid n_{ij} \in \text{Neighbors}(n_i)\}$
7:     $M \leftarrow \{m_{ij} \mid m_{ij} \in \{LF_m, HF_m\}\}$
8: **return** N, M

B. Learning Best Route

The protocol undergoes the learning phase where the nodes use a greedy action selection strategy to interact with the network and explore different routing paths. n_i selects a random one-hop neighbor as a relay for the first packet and observes the resulting network state (Lines 7 to 11 in Algorithm 2). Upon transmitting the packet, n_i calculates the Round-Trip Time (RTT), denoting the time a packet takes to travel from the source to the receiver and back. For an acoustic underwater channel, RTT is defined as:

$$\text{RTT}(s) = 2.1 \times \left(\frac{d_{ij}\,(\text{m})}{1500\,(\text{m/s})} \right), \qquad (14)$$

where $d_{ij}(m) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$ is the Euclidean distance between two neighbors exchanging packets, and 1500 (m/s) is the speed of sound underwater. As RTT ends, n_i checks for the reception of an acknowledgment (ACK) for the packet (Line 12 to 13 in Algorithm 2). If an ACK is received, the node updates the state transition probability $P^a_{s_i,s_j}$, computes the Q-value and V-value using equation 5, 11 and, 12, respectively (Lines 14 to 15 in Algorithm 2). Otherwise, packet p is retransmitted to the same node n_{ij} until the retransmission limit K is reached. Then, p is discarded (Lines 18 to 20 in Algorithm 2).

[1]We assume that all the nodes know their locations and the location of the sink node using the existing localization methods [4].

C. Adapting the Learned Information

In the adaptation phase, nodes incorporate the calculated Q-values of all their neighbor nodes to make routing decisions that achieve a trade-off between energy efficiency and reliability. This strategy balances exploring new routing paths and adapting the calculated Q-values to optimize future routing decisions in the adaptation phase (lines 22 to 24 in Algorithm 2). The protocol dynamically selects the appropriate modality and optimizes path selection to achieve energy efficiency by selecting nodes with the highest Q-value among the HF and LF neighbor nodes. REAM enables the nodes to refine their routing decisions continuously as the residual energy changes.

Algorithm 2 Find Best Relay and Mode

1: **Input:** α, γ, r_o, K, $E_{res}^{n_i}$, $E_{res}^{n_{ij}}$, Q_i , V_i.
2: **Output:** $a\,\langle n_{ij}, m_{ij}\rangle$
3: **for all** Packets $p \in \text{Queue}(n_i)$ **do**
4:     **if** sink $\in N$ **then**
5:         Send(p, sink) **continue**
6:     **while** $N \wedge M \neq \emptyset$ **do**
7:         **if** $p\,(n_{ij}) = 1$ **then**
8:             $a \leftarrow rand(n_{ij}, m_{ij})$
9:             $k \leftarrow 0$
10:             **while** $k < K$ **do**
11:                 Send(p, n_{ij}) $\wedge$ Set Timer to RTT
12:                 **if** Timer $= 0$ **then**
13:                     **if** Ack received **then**
14:                         Calculate $Q(s_i, a\,\langle n_{ij}, m_{ij}\rangle)$
15:                         Update $P^a_{s_i,s_j}$ and $V\,(s_i)$
16:                         **break**
17:                     **else**
18:                         $k \leftarrow k + 1$ & Repeat Steps 10 to 18
19:             **if** $k = K$ **then**
20:                 Discard packet
21:         **else**
22:             **for all** $n_{ij} \in N$ **do**
23:                 $a \leftarrow \arg\max_{(n_{ij}, m_{ij}) \in N \times M} Q(s_i, a\,\langle n_{ij}, m_{ij}\rangle)$
24:                 Repeat steps 9 to 20
25: **return** $a\,\langle n_{ij}, m_{ij}\rangle$

IV. Simulation Environment

To verify the effectiveness of our proposed routing method, we conducted extensive simulations using the UAN framework provided by Network Simulator 3 [5]. Our simulated environment replicates real underwater conditions relevant to our study using Thorp's model for acoustic absorption [6]. Our simulations are based on a Monte Carlo

TABLE I: Specifications of acoustic modems

Identifier	Modem	Range	Frequency	Data rate	Transmit Power	Receive Power
LF_m	S2CR 42/65	1000 m	42 KHz	31.2 kbps	18 W	0.8 W
HF_m	S2C T HS	300 m	120 KHz	62.5 kbps	3.5 W	0.8 W
SM	S2C R 18/34	3500 m	26 KHz	13.9 kbps	35 W	0.8 W

TABLE II: Simulation Parameters

Parameter	Value
UAN Deployment Space	$1000\,\text{m} \times 1000\,\text{m} \times 1000\,\text{m}$
Number of Nodes	5, 10, 20, 40, 60, 80, 100
Packet Inter-arrival Time (τ)	25, 50, 100 s
Time slot duration	20 ms
Packet size	40 B
Initial Energy (E_{ini})	1000 J
Simulation Time	3000 s
Discount Factor (γ)	0.5
Weight of energy cost (α)	1
Retransmission limit (K)	5

set of 95 CW sizes. Each simulation run consists of an underwater multi-modal acoustic network with 5 to 100 sensor nodes randomly deployed in a three-dimensional space of $1000\,\text{m} \times 1000\,\text{m} \times 1000\,\text{m}$. All the packets are routed to the sink node at the water surface through single and multi-hop transmissions. The sink node is capable of both acoustic and radio frequency communication and works as a gateway to the nearest base station as shown in Fig. 1. Each node is equipped with a set of Evologics acoustic modems [7] $M = \{LF_m, HF_m\}$, whose specifications are presented in Table I. Table II lists the simulation parameters and their values.

Packet Structure

The REAM protocol utilizes broadcast, data, and ACK packets to enable nodes to collect essential information for determining routing paths. Each packet consists of a 3-byte header and a payload of 30 bytes for broadcast and ACK packets and 40 bytes for data packets. Fig. 3 illustrates the information contained in each packet type.

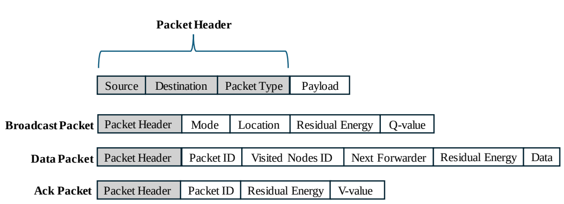


Fig. 3: Packet Structure

V. Results and Discussion

We consider a scenario where sensor nodes generate data traffic based on designated packet inter-arrival delay (τ). It determines the minimum time delay between individual packets to avoid forced collisions with a minimum of 23.3 sec (RTT $\times$ K) based on the longest-range acoustic modem identified as SM (see Table I for details). We consider $\tau = 100$ s, 50 s, and 25 s to indicate low, medium, and high traffic, respectively [8]. Nodes access the acoustic channel using CW-MAC protocol [9], with a random back-off window. For varying network density and traffic scenarios, Monte Carlo simulations were conducted to determine the optimal CW size that yields the maximum PDR from a set of values ranging from 5 to 100. The following results are obtained by averaging the acquired data over several simulations to achieve a confidence interval of 95%.

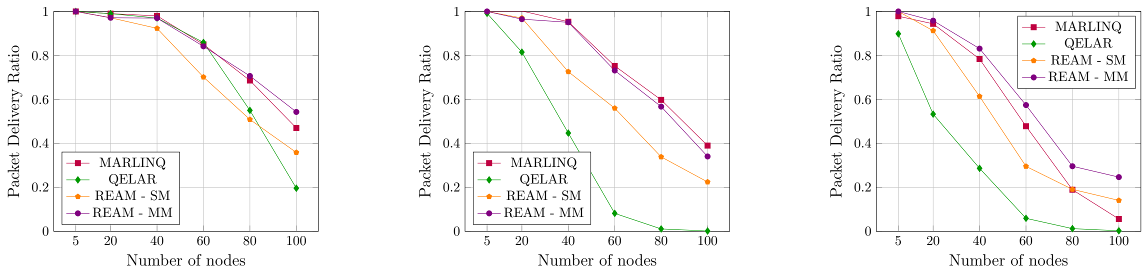
A. Performance Indicators

In this paper, we compare performance of the proposed REAM protocol with the following state-of-the-art protocols: (i) QELAR, a machine learning algorithm that minimizes and evenly distributes energy among groups of underwater nodes [10], and (ii) MARLIN-Q, a RL-based energy-efficient and latency-aware multimodal routing protocol [11]. We aim to validate the energy conservation at the bit level using multi-modal communication and to compare its performance through the following two performance indicators.

1) Packet delivery ratio (PDR): It refers to the ratio of the number of successfully received packets, $R_{p,i}$, to the total number of packets sent, $S_{p,i}$, in the network given as:

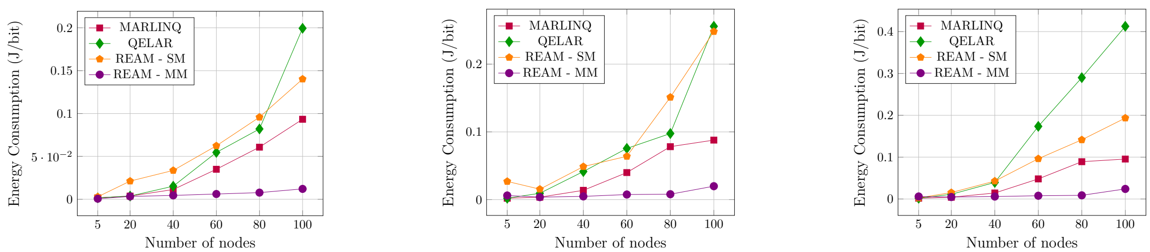
$$\text{PDR} = \frac{\sum_{i=1}^{n} R_{p,i}}{\sum_{i=1}^{n} S_{p,i}}. \tag{15}$$

Fig. 4 compares the PDR of REAM-MM, REAM-SM, QELAR, and MARLIN-Q for increasing network density and varying traffic. As network density increases, PDR declines across all algorithms due to higher contention, packet loss, and congestion in the acoustic channel. The results show that REAM-MM achieves comparable PDR to MARLIN-Q under low and medium traffic and outperforms it under high traffic, even when the network scales up. However, the improvement is significant compared to its single-mode variant, REAM-SM, and QELAR under all traffic conditions. This confirms that REAM-MM has robust mechanisms for managing network traffic by effectively handling packet collisions through retransmissions, optimal CW size, and RL-based efficient multi-modal routing strategies.



(a) Low Traffic ($\tau = 100\,\text{sec}$) (b) Medium Traffic ($\tau = 50\,\text{sec}$) (c) High Traffic ($\tau = 25\,\text{sec}$)

Fig. 4: Comparison of PDR for REAM vs. MARLIN-Q and QELAR under varying network load and traffic conditions.



(a) Low Traffic ($\tau = 100\,\text{sec}$) (b) Medium Traffic ($\tau = 50\,\text{sec}$) (c) High Traffic ($\tau = 25\,\text{sec}$)

Fig. 5: Comparison of energy consumption for REAM vs. MARLIN-Q and QELAR under varying network load and traffic conditions.

REAM-MM achieves higher PDR for high-traffic conditions than MARLIN-Q because it efficiently handles congestion by selecting the optimal CW size for different network loads. Additionally, for dense networks (above 40 nodes), when nodes get closer to each other, REAM-MM uses HF_m more frequently than LF_m, as depicted in Fig. 7. The availability of two modes creates a robust network of links as compared to MARLIN-Q whose routing mechanism chooses the modem based on incoming packet type. QELAR's low PDR is because it counterbalances energy consumption with routing efficiency. It chooses routes that are not the most reliable, particularly when node density increases, to distribute energy equally.

2) Energy consumption per bit: It quantifies the total amount of energy consumed by the network to successfully transfer a bit of data from the source to the sink. Suppose E_{net} represents the energy consumption of all the nodes and R_b is the total number of received bits at the sink then the energy consumed per received bit for the entire network is:

$$E_{R_b} = \frac{E_{\text{net}}}{R_b}. \tag{16}$$

Fig. 5 compares the energy efficiency[2] of the four algorithms for increasing number of nodes and three traffic scenarios. As the node density increases, the energy consumption per bit increases for all algorithms. This is anticipated as an increasing network density typically results in more routing overhead, collisions, retransmissions, and hop count, together consuming more energy. With a consistently comparable PDR to MARLIN-Q and a higher PDR than QELAR and REAM-SM, REAM-MM exhibits the lowest energy consumption, especially in dense networks. Fig. 5 shows that REAM-MM outperforms and achieves up to 81.2%, 79.6%, and 72.9% less energy consumption under low, medium, and high traffic scenarios, respectively, compared to the best-performing alternative, MARLIN-Q. The reason is threefold: (i) our protocol considers a node's residual energy in the cost function (Section II-A4), and (ii) the use of HF_m that requires less power for communication, thereby minimizing per-bit energy consumption, and (iii) the use of the most fitting CW size. The right CW size saves unnecessary retransmissions by controlling collisions while maintaining reliable packet delivery. MARLIN-Q,

[2]These results contain the energy consumption in phases 2 (section III-B) and 3 (section III-C). Due to page limitations, the average energy consumption of a node in all three phases is not included.

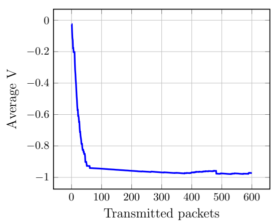


Fig. 6: Average $V^{\pi}(s)$ of all the nodes.

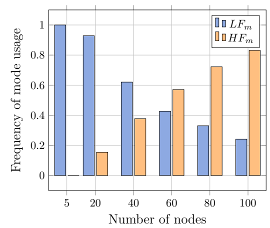


Fig. 7: Frequency of modes in varying network load.

REAM-SM, and QELAR show relatively higher energy consumption than REAM-MM across dense networks. It is because MARLIN-Q employs a cost function that specifically considers latency, thus optimizing end-to-end delay at the cost of higher energy consumption. Conversely, QELAR's energy efficiency is lower due to its reliance on a single long-range transmission mode, high computational overhead from persistent channel listening, and energy costs from periodic broadcast control packets to update the Q-table.

B. *REAM Convergence*

Fig. 6 shows the efficiency of the proposed REAM protocol in terms of the average $V^{\pi}(s)$ with a network size of 100 nodes transmitting packets under high data traffic, i.e., $\tau = 25$ s. Initially, the algorithm explores the state-action space with minimal prior knowledge, trying various actions to learn the network. This results in a continuous decay in V values, indicating the learning phase. After 50 packets, the rate of change in Q-values slows down as the algorithm starts exploiting learned actions, selecting minimal energy paths that generate high rewards. The graph stabilizes, showing little to no change in average V values suggesting that the algorithm has reached $\pi(s)$ by consistently opting for actions that minimize cumulative energy costs. The stabilization rate reflects the convergence speed to reach optimal policy and the effectiveness in achieving its goal of energy-efficient routing.

VI. Conclusion

REAM is a fully distributed and self-adaptive Q-Learning-based multi-modal underwater routing protocol that facilitates the transmission of data from underwater sensors, instruments, and AUVs to the surface while enhancing reliability and minimizing energy consumption in UASNs. By providing connectivity and enabling energy-efficient data exchange, REAM supports collecting scientific data from underwater sensors while effectively handling the underwater network dynamics. Results show that our proposed framework, REAM-MM, achieves lower energy consumption with a comparable PDR under low and medium traffic, and relatively higher PDR in high-traffic scenarios compared to the best-performing alternative, MARLIN-Q.

References

[1] A. Pal, F. Campagnaro, K. Ashraf, M. R. Rahman, A. Ashok, and H. Guo, “Communication for underwater sensor networks: A comprehensive summary,” *ACM Transactions on Sensor Networks*, 2022.

[2] E.-C. Liou, C.-C. Kao, C.-H. Chang, Y.-S. Lin, and C.-J. Huang, “Internet of underwater things: Challenges and routing protocols,” in *2018 IEEE International Conference on Applied System Invention (ICASI)*, 2018.

[3] M. L. Puterman, “Markov decision processes,” *Handbooks in operations research and management science*, 1990.

[4] H.-P. Tan, R. Diamant, W. K. Seah, and M. Waldmeyer, “A survey of techniques and challenges in underwater localization,” *Ocean Engineering*, vol. 38, no. 14-15, pp. 1663–1676, 2011.

[5] “Uan framework — model library.” https://www.nsnam.org/docs/models/html/uan.html. (Accessed January 14, 2025).

[6] W. H. Thorp, “Analytic description of the low-frequency attenuation coefficient,” *The Journal of the Acoustical Society of America*, 1967.

[7] “Homepage-evologics.” https://evologics.com/. (Accessed January 14, 2025).

[8] S. Basagni, V. D. Valerio, P. Gjanci, and C. Petrioli, “Harnessing hydro: Harvesting-aware data routing for underwater wireless sensor networks,” *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2018.

[9] N. Parrish, L. Tracy, S. Roy, P. Arabshahi, and W. L. Fox, “System design considerations for undersea networks: Link and multiple access protocols,” *IEEE Journal on Selected Areas in Communications*, 2008.

[10] T. Hu and Y. Fei, “Qelar: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks,” *IEEE transactions on mobile computing*, 2010.

[11] S. Basagni, V. Di Valerio, P. Gjanci, and C. Petrioli, “Marlin-q: Multi-modal communications for reliable and low-latency underwater data delivery,” *Ad Hoc Networks*, 2019.