Article

PACKET FLOW BASED REINFORCEMENT LEARNING MAC PROTOCOL FOR UNDERWATER ACOUSTIC SENSOR NETWORKS

Ibrahim B. Alhassan [1,†,‡] 0000-0002-2125-2034, **and Paul D. Mitchell** [1,‡] 0000-0003-0714-2581

[1] Department of Electronic Engineering, University of York, York UK 1; ibrahim.alhassan@york.ac.uk
[2] Department of Electronic Engineering, University of York, York UK 2; paul.mitchell@york.ac.uk
* Correspondence: I,A., ibrahim.alhassan@york.ac.uk; P.M, paul.mitchell@york.ac.uk

Abstract: Medium access control (MAC) is one of the key requirements in underwater acoustic sensor networks (UASNs). For a MAC protocol to provide its basic function of efficient sharing of channel access, the highly dynamic underwater environment demands MAC protocols to be adaptive as well. Q-learning is one of the promising techniques employed in intelligent MAC protocol solutions, however, due to the long propagation delay, the performance of this approach is severely limited by reliance on an explicit reward signal to function. In this paper, we propose a restructured and a modified two stage Q-learning process to extract an implicit reward signal for a novel MAC protocol: Underwater packet flow Aloha with Q-learning (ALOHA-QUPAF). Based on a simulated pipeline monitoring chain network, results show that the protocol outperforms both Aloha-Q and framed Aloha by at least 13% and 148% in all simulated scenarios respectively.

Keywords: MAC protocols; Reinforcement Learning; Underwater Acoustic Sensor Networks

1. Introduction

Medium access control (MAC) is one of the key requirements in underwater acoustic sensor networks (UASNs) garnering a major interest in the research community [1–3]. As an analogue of terrestrial sensor networks, UASNs are envisaged to enable a multitude of civilian and military applications [4–6]. To advance these applications, sensor nodes are being developed to be small/compact for easy transport, given the environment is characteristically challenging to access. There is interest in new sensor nodes being energy efficient for longer deployments as currently there is no viable energy harvesting technology, they should also be and cheap to lower the overall cost, since, UASNs are envisaged to be deployed to cover substantial marine areas and require a large number of devices. Employing acoustic waves in UASNs imposes some channel-centred unique constraints, such as: limited distance and frequency dependent capacity (bandwidth and data rate), long and variable propagation delay, and high bit error rate (BER) on the design of UASNs [2,4,7]. As such, there is growing demand for efficient MAC solutions, especially adaptive MAC protocols for practical networks in the highly dynamic underwater environment.

Although preliminary studies on adopting existing MAC techniques/schemes from the vast body of work on terrestrial MAC protocols to underwater networks was largely found to be ineffective [1,8], the insight from the underlying principles remains useful. As a general guide, the network topology gives an insight into the appropriate category of MAC scheme to employ, with contention-free and contention-based schemes better suited to centralised and decentralised topologies respectively. Centralised topologies typically facilitate schedule creation and coordination from a central controlling node. Therefore, uncoordinated channel access becomes too contentious and less efficient. On the other hand, in a decentralised topology, such coordination is prohibitively challenging to implement, and the limited resources makes contention-free protocols inefficient.

Citation: Alhassan, I. B; Mitchell, P. PACKET FLOW BASED REINFORCEMENT LEARNING MAC PROTOCOL FOR UNDERWATER ACOUSTIC SENSOR NETWORKS. *Sensors* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Code Division Multiple Access (CDMA) and Frequency Division Multiple Access (FDMA) are promising contention-free schemes considered for UWASNs [9,10]. CDMA assigns unique binary codes to users (nodes) to spread the information signal thereby offering the complete frequency band to nodes for simultaneous transmissions. Frequency hopping and direct sequence spread spectrum (FHSS and DSSS respectively) are the standard modulations employed in this scheme. FDMA splits the channel into distinctive frequency bands and assign them to different users. In this way, users can initiate concurrent transmissions without incurring collisions [5,10]. While the radio bandwidth (GHz) enables implementation of these schemes with relative ease, in UANS the available bandwidth is very limited (KHz).

Time Division Multiple Access (TDMA) [11] creates schedules by splitting time into slots, and is the most promising contention-free approach used in UASNs, because of its flexibility and potential to achieve true collision free scheduling. Despite the challenges of synchronisation, some solutions leverage the long propagation delays for spatial reuse to improve performance. A gateway node in [12] creates a gap-free schedule, and then requests packets from the transmitting nodes. Other solutions incorporate sleep-cycles between activity to save energy [3]. The solution in [13] is for a central node to use an initialisation stage to gather network wide information that is then optimised using Genetic and Particle Swarm Algorithms to create a collision free schedule. However, the lack of complete knowledge of the environment poses a major challenge for creating a lasting collision-free schedule.

Contention-based MAC protocols such as Aloha [14] and its variants offer low complexity and simplicity of implementation. The downside is that contention-based protocols suffer low utilization and prohibitively large end-to-end delay at high loads due to the blind transmission strategy. The works in [15,16] integrate additional guard time between successive transmissions in order to reduce collisions, and [17] demonstrates receiver initiation (RI) to improve the performance. In RI, the receiver makes the first move of initiating the data transfer session by sending a request packet to the transmitter(s) (essentially polling). Since collision occurs at the receiver, the RI approach aims to eliminate the most common source of collision: transmit-receive collision. All these approaches add to the complexity, and the overheads incurred by the control packets limit the achievable utilisation.

A popular technique is to incorporate both contention-based and contention-free components to form hybrid MAC protocols. This strategy improves performance by allowing networks/devices to switch to an optimum MAC scheme based on demand or traffic profiles. Variations in traffic is addressed in [18] where the protocol is pre-configured to assign capacity either by free assignment or on demand, and [19] balances performance with two time slots in a frame, one slot for scheduled transmissions and the other for random access.

In the highly dynamic underwater environment, MAC protocols need to be adaptive to changing conditions as well. This is because previous assumptions used to create schedules may be outdated or sub-optimal due to changes in topology, traffic, node(s) failure(s) and/or addition(s). Reinforcement learning is a promising solution used in MAC protocols to provide adaptability and robustness in wireless sensor networks, such as ad-hoc emergency networks for disaster monitoring [20,21]. In such networks, intelligent MAC protocols will adapt to the changing topology or the environment. Instead of switching between MAC schemes, reinforcement learning is used to continually assess the network condition through feedback and with a view to maintaining (in so far as possible) a collision free schedule.

In [21] we studied the use of Aloha-Q [20] underwater. Aloha-Q is a MAC protocol originally developed for terrestrial wireless sensor networks. It employs a Q-learning algorithm to incorporate intelligence into framed-Aloha. The frame is created with a predetermined number of periodic time slots. Each slot is structured such that it accommodates a data packet, an ACK packet and their corresponding 1-hop propagation delays (Figure 3). Initially, nodes randomly select and transmit in any slot, but eventually each node settles on a collision-free slot as the underlying Q-learning reward/punishment serves to reinforce successful slots. However, because the ACK serves as the critical signal for the reward/punish mechanism in the Q-learning algorithm, the overhead with respect to the slot size due to the long propagation delay severely constraints the effectiveness of the Q-learning strategy in terms of achievable utilization and end-to-end delay.

In section 3.2 we demonstrate the Q-learning update mechanism, and how it is applied in the Aloha-Q protocol.

The focus of this paper is to implement a robust, simple and computationally inexpensive MAC protocol that consistently and efficiently delivers the maximum channel utilisation in a monitoring chain UASN, such as in underwater pipeline monitoring. To achieve that, we are inspired by the research in [20,22–24].

Specific contributions are:

- Provide some background work on the feasibility of restructuring the slot size in a typical frame-based underwater MAC protocols to improve network performance.
- To propose a new slot structure with minimal overhead based on the relationship between packet transmission duration and the 1- hop propagation delay that is capable of achieving the theoretical channel utilization.
- To propose Aloha-QUPAF, a novel dual-control intelligent approach to medium access control based on packet(s) flow in a linear chain network.

The rest of the paper is structured as follows. Section 2 introduces the frame-based approach of mac protocol design and the network model. Section 3 presents the proposed slot size and the analytical modelling. Section 3.1 discusses the simulation results based on the network model as compared to the theoretical results. It is followed by section 4, our detailed dual-control intelligent MAC scheme, and the results obtained when applied to varying lengths of chain networks. Finally, we draw conclusions.

Table 1: Table of Mathematical Terms

Entry	Description
N	Number of nodes
S_L	Number of slots per frame
N_{opt}	Optimum number of slots per frame
U	Channel utilisation
τ_d	Data packet duration
τ_A	ACK packet duration
τ_g	Guard duration
$K\tau$	ratio τ_d to τ_{pg}
S^a	Slot size with ACK
S^n	Slot size without ACK
α and γ	Learning rates
λ	Optimisation scale
fl_τ	Packet flow average

2. Frame-Based MAC Protocol

In this section, an overview is given on the fundamental operation of a baseline frame-based random access protocol. With the aid of a simple network model we analyse and identify the limitations of frame-based scheduling (in terms of achievable channel utilization) with a random access scheme.

Framed-Aloha is one of the baseline protocols we compare against our proposed intelligent scheme. In contrast to slotted Aloha, whereby time is divided into slots and nodes can only transmit at the beginning of each slot, a frame is used in framed-Aloha, which comprises a fixed N_s number of contiguous slots. In the framed-Aloha random access strategy each node independently and randomly chooses one of the transmission slots at the beginning of each frame.

Typically, a slot is structured such that it accommodates: a data packet of duration (τ_d), acknowledgment packet of duration (τ_A *if required*), the associated propagation delays of each packet (τ_{pg}), and a small guard band(τ_g): the band is essential to correct and guard against drifts in clock precision and synchronisation. The slot structure is shown in Figure 1, for cases with and without acknowledgments. Whereas, in radio networks, the overheads due to the wait period between successive data transmissions in a slot/frame can be of negligible length with respect to the packet duration, in an underwater acoustic channel, however, the physics impose a long propagation delay, plus low capacity, (bandwidth and therefore data rate) making the overheads significant, thus, negatively impacting the channel utilization and end-to-end delay.

Defining the channel utilization (U) as the rate of delivering data at the designated sink node (Equation 1), then, in frame/slot based protocols, the utilization is also a function of the number of slots (N_s) in the frame. For example, if a node is allowed to transmit N packets per frame, then the maximum effective utilization at the sink is going to be upper bounded at N/N_s. The value of N_s is determined from the topology and interference population of the network. Setting N_s inappropriately will negatively affect not just the utilisation but potentially the stability of the MAC protocol as well. For example, in a star topology N_s is equal to the number of transmitting nodes (N_n), as each node should have a unique transmitting slot, setting $N_s > N_n$ adds extra un-utilised slot(s), and $N_s < N_n$ will cause contention as some nodes will not exclusively own a slot. Therefore, for a particular topology and an interference model, there is an optimum N_s (N_{opt}) [20]. Erlang [25] is a dimensionless unit that represents continuous channel usage (for example 0E = zero channel activity, 0.5E = half channel activity and 1E = full channel usage).

$$U_{normalised}(Erlang) = \frac{N \times \tau_d(s)}{N_s \times S(s)} \tag{1}$$

therefore, the optimum Utilization is:

$$U_{normalised}(Erlang) = \frac{N \times \tau_d(s)}{N_{opt} \times S(s)} \tag{2}$$

where S is the slot duration.

One of the consequences of having low capacity is the long transmission duration, which presents two situations for a given transmitter and receiver pair; the transmission duration is either greater than or less than the propagation delay between the nodes. Following [26], if we introduce the parameter K_τ (Equation 3), then, the resulting slot structure can have either of two sets of transmission-reception patterns: overlapping and non-overlapping based on the value of $K\tau$, as shown in Figure 1.

$$K\tau = \frac{\tau_d}{\tau_{pg}} \tag{3}$$

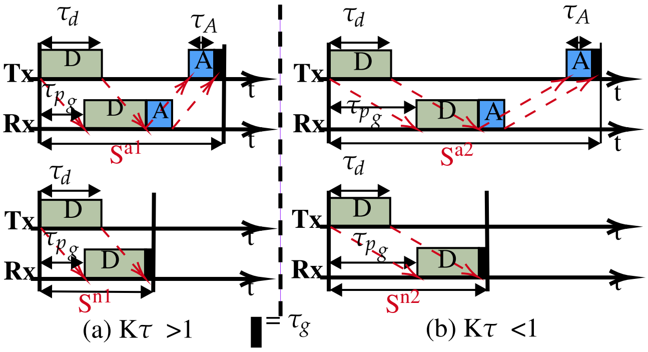


Figure 1. Typical slot structures: (a) Overlapping, transmission-reception occurs concurrently for data packet (b) Non Overlapping, data transmission completed before reception occurs

S^{a1} and S^{a2} represent the slots length with ACK and are typically used by slotted protocols employing an ACK signal such as ALOHA-Q. Similarly, S^{n1} and S^{n2} are the slots without ACK as used in framed-Aloha and TDMA. Equations (4 and 5) are used to calculate the slot sizes.

$$S^a = \tau_d + \tau_A + 2\,\tau_{pg} + \tau_g \tag{4}$$

$$S^n = \tau_d + \tau_{pg} + \tau_g \tag{5}$$

In this slotted concept, nodes are allowed to transmit only one packet per frame (i.e N=1) and the expression of maximum utilisation (U) can be simplified to the ratio of packet duration to frame size (Equation 6). We can combine Equation 2 and Equation 6 to calculate the expression of the utilisation below;

$$U = \begin{cases} \frac{\tau_d}{N_{op_t}\ (\ \tau_d + 2\tau_{pg} + \tau_A + \tau_g\ )}, & S^a \\ \frac{\tau_d}{N_{op_t}\ (\tau_d + \tau_{pg} + \tau_g\ )}, & S^n \end{cases} \tag{6}$$

as $\tau_d, \tau_{pg} >> \tau_A, \tau_g$, Equation 6 approximates to;

$$U \approx \begin{cases} \frac{\tau_d}{N_{op_t}\ (\ \tau_d + 2\tau_{pg}\ )}, & S^a \\ \frac{\tau_d}{N_{op_t}\ (\tau_d + \tau_{pg}\ )}, & S^n \end{cases} \tag{7}$$

From Equation 7, it can be seen that, since τ_d and τ_{pg} dominate, the value of $K\tau$ will guide us on how to improve channel utilisation by restructuring the slot size. For $K\tau > 1$ we are constrained with respect to any change to the slot size. Any reduction will create overlapping slot reception that will effectively render the slotting meaningless, as demonstrated with the downgrade of slotted Aloha to pure Aloha underwater [26].

In most UASNs applications, the propagation delay is longer than the transmission duration because of sparse connectivity. Therefore, $K\tau < 1$ best describes such scenarios. We propose the slot structure in Figure 3. The slot size is now reduced to approximate the propagation delay ($S \approx \tau_{pg}$), which is possible since with $K\tau < 1$ the data packet can be safely accommodated in τ_{pg}. This simple slot structure aims to reduce and fill the otherwise wide gap in the conventional slots with useful data (compared to Figure 1). Therefore, for a given chain UASN, designed with nodes separated by dm transmission range, we will demonstrate that there are advantages in performance improvements of using our slot structure. For example, the peculiar characteristic of the underwater communication channel in terms of its distance-dependent capacity, that is, acoustic transmission bandwidth and data rates decrease with increasing transmission distance [27]. As such, instead of few hops transmitting over longer ranges (requiring high power) with low capacity, we can potentially achieve higher capacity transmissions with additional hops added to route data over shorter ranges (low power). To investigate the achievable utilisation, the slot structure shown in Figure 3 is based on $K\tau \approx 1$: a special case of $K\tau < 1$. This is purely to limit the overhead in the slot, as increasing the slot size beyond τ_{pg} negatively affects the utilisation according to Equation 6.

2.1. Scenario and Network Model

Consider a scenario comprising quasi-stationary equally spaced nodes in an N-hop underwater network chain topology, with data delivered along the chain from one end to the other. Figure 2, depicts an example of such network with N = 4, and hop distance d. This topology is representative of pipeline monitoring. As such, during the reporting cycle, the network can be considered loaded to capacity, accordingly this work is primarily concerned with the achievable utilisation. To aid the analysis, the following assumptions are made;

1. All nodes are homogeneous and communicate over a single channel, half-duplex mode.
2. Collision model (non capture) is used, i.e if two or more packets overlap at the receiver they are discarded.
3. Nodes are globally synchronised, an assumption commonly employed to simplify analysis and applicable to quasi-stationary nodes synchronised before deployment.
4. Interference range (Ifx) is twice the reception range (Rx), this model is typically employed for chain networks as an illustrative model to incorporate the effect of interference from nodes that are 2-hop away .
5. A source node has saturated traffic, i.e always has a packet to send, to provide the maximum monitoring rate based on the transmission opportunities offered by the MAC layer. Similar research papers are concerned with achievable utilization [23,28,29].

6. All source/relay nodes can only transmit one packet per frame, a consequence of assumption (4) yields a frame consisting of four slots [20], as only 1-of-4 connected nodes can transmit successfully at a given time.

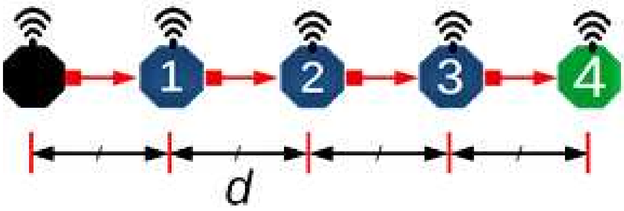
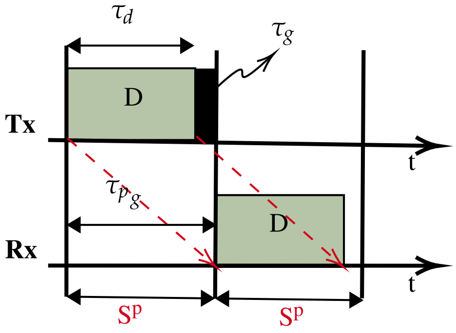


Figure 2. An example scenario



(b) Kτ ≊1

Figure 3. Proposed Slot Structure

We re-write Equation 7 of S^n to get the new utilisation for the proposed slot structure;

$$U_{normalised}(Erlang) = \frac{\tau_d}{N_{op_t} \times \tau_{pg}} \tag{8}$$

and in terms of K_τ becomes;

$$U_{normalised}(Erlang) = \frac{K_\tau}{N_{op_t}} \tag{9}$$

In summary, while the traditional slot structure that incorporates the propagation delay and/or ACK packet within constraints of the available channel resources, we showed that with $K\tau < 1$ the propagation delay is sufficient to accommodate data packet, then it is possible for the slot size to be effectively reduced and restructured (by at least 50% of the cases in the $K\tau < 1$ regime) and as long as a protocol does not requires an Ack packet, there is a potential for a dramatic improvement in performance (Equation 9 vs Equation 7).

3. Model Analysis

To analyse the network with the proposed slot structure (Figure 3), we consider a baseline scheme whereby each node initialises by randomly choosing a transmission slot. The purpose of considering this scheme is firstly, to demonstrate the inefficiency of a random access scheme by analysing the distribution of the achievable channel utilization. Secondly, to investigate the feasibility of applying intelligent techniques to the model that could lead to a significant performance improvement. Finally, to evaluate the efficacy of the proposed slot structure coupled with the intelligent techniques relative to similar intelligent approaches and random access baseline schemes.

To build the frame, we start with the optimal number of slots per frame N_{opt}, in a linear chain network (such as Figure 2 and longer) N_{opt} is four as computed according to the 2 hop interference model [20]. This is because in a linear topology with 2 hop interference model, technically only one in four nodes can successfully transmits at a given time. Similarly, for 1 hop and 3 hop interference models one in three and one in five nodes could transmit successfully [20,23]. Therefore, for a distributed MAC protocol, such as framed-Aloha employed in this setup, each node is free to chose any of the available four slots in the frame resulting in $4^4 = 256$ ways for nodes to independently select and occupy transmission slots. Table 2 lists the range of the 256 possible slot combinations in a four column array of 64 unique patterns, with each column vector signifying the transmission slot pattern from node 0 to node 3. That is, the vector $[0000]$ denotes all nodes selecting and occupying slot 0, likewise, slot sequence $[2210]$ signifies both nodes 0 and 1 choosing slot 2 while nodes 2 and 3 choose slots 1 and slot 0 respectively. Pictorial timing depictions (see appendix A) are employed to observe and obtain the theoretical bounds of the scheme in terms of channel utilisation. The diagrammatic method provide a visual intuition of our core idea. In appendix A, six examples (Figure A2 to Figure A7) are provided to illustrate the process. For each pattern, N_0 is the source node, it generates and transmits data in every frame to N_1 which forwards the packet (if successfully received) to N_2 in the next frame and so on. Overall, individual packets are traced frame-by-frame as they traverse the network from source to sink (N_0 to N_4). The final utilisation is measured when an overall periodic pattern emerges at the sink node (vertical red lines in each example figures, refer appendix A).

Table 2: Possible Slot Permutations

S/N	**SLOT SEQUENCE**			
	SEQ_0XXX	SEQ_1XXX	SEQ_2XXX	SEQ_3XXX
0	[0 0 0 0]	[1 0 0 0]	[2 0 0 0]	[3 0 0 0]
1	[0 0 0 1]	[1 0 0 1]	[2 0 0 1]	[3 0 0 1]
...	[...]	[...]	[...]	[...]
...	[...]	[...]	[...]	[...]
62	[0 3 3 2]	[1 3 3 2]	[2 3 3 2]	[3 3 3 2]
63	[0 3 3 3]	[1 3 3 3]	[2 3 3 3]	[3 3 3 3]

3.1. Results

In order to empirically evaluate the performance of above random access scheme, we run a simulation on a network of 5 nodes (Figure 2) configured with the proposed slot structure analysed in Section 3. Each node is pre-configured to run a MAC protocol that randomly selects and maintains a transmission slot at the beginning of each simulation run. It should be noted that in this simulation, since $K_\tau \approx 1$, the transmission delay and propagation delay are abstracted to 1 : 1 for best results.

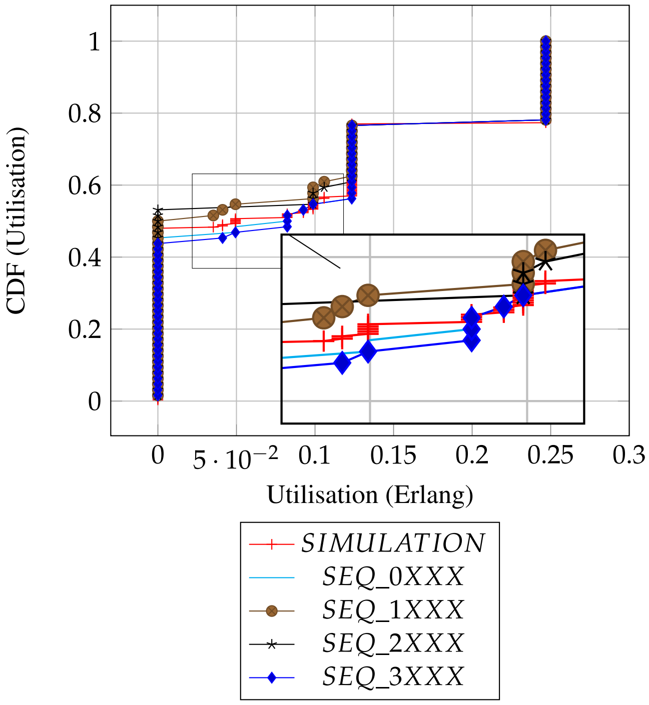


Figure 4. Distributions Comparison

Figure 4 shows and compares the utilisation results from both the analytical distributions of the slot patterns and the simulations. Overall, there are three dominant utilisation levels and some spurious intermediate levels as summarised in Table 3. The summary provides individual proportions of levels in each plot, and the overall column is the contribution of each sequence in the combined set of 256 slots.

Table 3: Summary of Utilisation Levels

Level	**Proportions(%)**				
	SEQ_0XXX	SEQ_1XXX	SEQ_2XXX	SEQ_3XXX	Overall
Worst case (0E)	45.3	50.0	53.1	43.8	48.1
Intermediate (0.03E - 0.1E)	9.4	10.4	6.3	10.9	9.4
Half (0.125E)	21.9	15.6	17.2	21.8	19.1
Maximum (0.25E)	23.4	23.4	23.4	23.4	23.4

Depending on the chosen slot by the source node, transmissions could be initiated from either the frame edge (slots 0 or 3) or mid-frame (slots 1 or 2) and to some degree, the results show how the position of a chosen slot affects the utilisation. As shown in the result summary (Table 3), there is a subtle but clear advantage in performance when the source node initiates transmissions with emerging slot patterns at frame edges (i.e **SEQ_0XXX, SEQ_3XXX**) relative to the mid frames (i.e **SEQ_1XXX, SEQ_2XXX**) or there is at least 8% better chance of getting a packet received at the sink node when the source node transmits at the edges of a frame as compared to when source node use mid frame (in terms of the worst case utilisation levels).

Intuitively, the distribution of the utilisation of the patterns can be assumed to be similar, since it can be demonstrated that each column sequence can be translated to another corresponding sequence in the remainder of the columns (Table 2). However, due to the transmission strategy of the protocol of scheduling packet transmission at the beginning of each frame, the simple slot structure guarantees that packets transmitted at $slot_i$ be received at $slot_{i+1}$. This means sequence translations will result in packet reception/interference across frames, consequently causing the distribution of the utilization outcomes to vary. For example, consider the corresponding slot selection sequences: [0 0 3 0], [1 1 0 1], [2 2 1 2] and [3 3 2 3]. [0 0 3 0] and [3 3 2 3] both have cross-frame receptions and have a similar utilization of 0.125 Erlangs (Figure A7). In contrast [1 1 0 1] and [2 2 1 2] have no cross frame reception and yield 0 Erlangs (appendix A:

Figure A2). Only 60 out of the total 256 slot sequences yield the maximum utilization level as a whole, and remain immune to the slot sequence translations because, they are perfectly collision free. In Figure A2 to Figure A7 (appendix A), we show how we computed six (6) of the ten (10) prominent utilisation levels for brevity.

The simulation results are in agreement with our analytical results, as they show that no data gets delivered 48% of the time. This corresponds to the average of the possible 43%-53% worst cases in the given original slot patterns as expected. Most importantly, the simulation result confirms that the full channel utilization is achievable with the exact proportion of 23%. Finally, the simulation result shows the average performance of the random slot selection protocol and will serve as a baseline with which to demonstrate the merit of slot based learning in the new protocol ALOHA-QUPAF.

3.2. Q-Learning

This section demonstrates the underlying Q-learning update procedure based on stateless Q-learning [30]. Following the standard Q-learning framework, an agent learns how to behave in an unknown environment by interaction with the environment. The agent perceives and changes the state of the environment by taking an action, receives a response from the environment which indicates the quality of the action taken in form of a reward/punish signal. This process is Markovian and it is modeled as an MDP [30–32]. To develop a MAC protocol, this is translated to node taking the action of transmitting data packet and the successful/unsuccessful reception of an ACK packet represents the reward/punish signal. Each node is given a vector of Q-values (Q-table), each Q-value is in turn assigned to one slot in the frame (section 1). At the beginning of each frame, a node will scan the Q-table and select the slot with the highest Q-value to schedule transmission in that slot. Successful transmissions are rewarded and unsuccessful transmissions are punished based on the reception or otherwise of an Ack packet and updating the Q-value of the transmission slot using Equation 10.

$$Q[S_t] \leftarrow Q[S_t] + \alpha(\psi - Q[S_t]) \tag{10}$$

Where $Q[S_t]$,α and ψ denote the Q value of the current slot, the learning rate (0.1) and the reward/punish signal(± 1)

Table 4 illustrates an example of the Q-learning as implemented in Aloha-Q. Consider an initial situation (Frame 0) whereby a node i with data to send randomly chooses slot 2 (because all slots have equal Q-values) at the beginning of a frame to schedule transmission and the transmission was unsuccessful.

- The new Q-value of slot 2 becomes;
  $Q[2] \leftarrow 0 + 0.1(-1 - 0)$; [-0.1]
- In the next frame slot 2 has the lowest Q-value and is not considered, the node again chooses slot 1 randomly (among slots 0, 1 and 3). Following a successful Ack reception, the new Q-value of slot 1 is updated.
  $Q[1] \leftarrow 0 + 0.1(+1 - 0)$; [0.1]
- Frame 2, the node chooses slot 1 as it has the highest Q-value (0.1) and send data, with successful Ack reception the Q-value is updated accordingly.
  $Q[1] \leftarrow 0.1 + 0.1(+1 - 0.1)$; [0.19]

The table gives the Q-values up to twenty frames assuming slot 1 continues to be successful. This simple, yet effective recursive Q-learning update bootstraps the trial-and error mechanism to a robust collision-free schedule as each node will eventually and independently occupy a unique transmission slot.

Table 4: Example of Q-value update in Aloha-Q

Frame/Q-values	Q[0]	Q[1]	Q[2]	Q[3]
FRAME 0	0	0	0	0
FRAME 1	0	0	-0.1	0
FRAME 2	0	0.1	-0.1	0
FRAME 3	0	0.1900	-0.1	0
FRAME 4	0	0.2710	-0.1	0
...	...	...	...	...
FRAME 20	...	0.8499	-0.1	0

However, as previously stated, while the Ack signal is crucial to the Q-value update operation, it puts an additional burden on the scarce network resources underwater: reducing utilisation due to overheads and increased delay due to the Ack signal wait times. Our goal is to implement a novel Q-learning approach that maintains the level of intelligence without this explicit Ack signal, thereby maximising the channel utilisation and improving end-to-end delay.

4. Underwater Packet Flow Aloha-Q: Aloha-QUPAF

The proposed slot structures in Figure 3 pose a critical question: how do we apply a simple reinforcement learning algorithm to ultimately achieve collision free scheduling without an ACK packet? In this section we present a two stage solution using a reformulated Q-learning coupled with a simple stochastic averaging expression [33]. We will demonstrate the efficacy of our dual-mode learning approach in improving performance in a chain network as introduced in section 2.1.

4.1. Protocol design

In order to achieve the goal of realising a collision-free schedule without an explicit Ack signal, we modified the Q-value update process (section 3.2) while maintaining the remaining protocol settings and assumptions (section 2.1 and section 3.2). Specifically, at the beginning of each frame a relay node chooses the slot with the highest Q-value (if more than one slot has highest Q-values, one is chosen at random) to forward a received packet on to the next hop. In the case of the source node, it initialises by randomly selecting and maintaining a constant slot for transmission. This is because we employ a Q-learning process that utilises packet receptions to update and reinforce transmission slot selection. Our solution involves a two stage approach based on the following intuitions:

1. In a network with half-duplex nodes, they cannot transmit and receive at the same time (slot), therefore we employ Q-learning to isolate all reception slots by punishing those slots to lower their Q-values. As such, when a node scan the Q-table, receptions slots will have low Q-values and are unlikely to be selected for transmission.
2. A continuous flow of packet(s) over the chain is expected in a saturated traffic with a healthy channel. Thus, a relay/sink expects new packet(s) in every frame after receiving the first packet and a packet collision is inferred whenever that stream of packet(s) gets disrupted. To exploit this realisation, every time a relay node transmits a packet it rewards the chosen transmission slot (positively updates the slot's Q-value) if and only if a new packet is received afterwards.

We denote the two stages in the dual mode control as slot selection and flow harmony and a detailed description of the process is given below:

- **Slot selection:** Is implemented by Q-learning to eliminate reception slot(s). When a source node generates a packet and transmits, upon receiving the packet, the receiver (relay node) will record the reception slot (rx_s) and update the Q value of the slot according to (Equation 10). Specifically, each slot in a frame is mapped to a value in the vector of $\mathbf{Q}$ values ($Q[ns]$), the Q values are initialised with a uniform random number less than 1, whereby for each reception, the node computes rx_s and updates $Q[rx\_s]$ accordingly with $\psi = -1$. Consequently, this continual negative reinforcement of reception slots isolates those slots and the slot(s) with the highest Q value signifies a probable collision free slot at the local level, therefore good candidate(s) slot(s) for transmission. For a relay node, at the beginning of each frame, if a node has packet(s) in its queue, it will schedule a packet transmission in a slot with the maximum Q value, however if more than one slot shares the maximum Q value, one will be chosen at random from amongst them. Whilst the Q-value of the reception slot is always punished following any reception, the Q-value of the transmission slot is only updated after every transmission. If there is a subsequent packet reception, the transmission slot is rewarded ($\psi = 1$) otherwise it is punished ($\psi = -1$). However, since this scheme lacks a definitive feedback signal based on this node action(s) of transmissions, the success of any transmission in the chosen slot is uncertain. This is because, unless if the packet flow is network wide, a continuous transmissions and reception by a relay node does not mean that a given node's transmissions are not interfering with some other transmissions especially for the downstream links. Therefore, to avoid nodes from getting stuck in a local minima, a control mechanism has to be devised to regulate the Q values especially of the transmission slot.

- **Flow harmony:** Although we devise a means to obtain feedback from the environment (reward/punishment), the node cannot directly link these signals to its own action(s), hence, at any given time during the network run, we only have a partial observation of the channel condition, this type of process is best modeled as a Partially Observable Markov Decision Process (POMDP) [34,35]. This is because, instead of a certainty in the network wide flow, the packet flow experienced by each node gives us a partial observation on the channel at the local level. The POMDP framework enable us to model the local observations by agents to generate a probability distribution of a belief state (in our case settled or unsettled flow). The network can be in either stable or unstable packet flow states, we therefore designate two belief states accordingly. We employ a simple heuristic strategy based on the stochastic averaging [36] whereby each node independently tracks its overall local packet flow in a given window, which we then translate as the distribution of the belief state. The distribution of the belief states is computed with Equation 11. For each reception in a frame, fl_τ is updated by λ steps at the tracking rate γ. While the expression monotonically approaches 1, it is continually windowed every (W_n) frames and compared to a fixed threshold (*thresh*). Based on our simulation experiment, ideally fl_τ will reached 98% by the 20^{th} frame, hence, we heuristically set ($W_n = 20$) to check for fl_τ with a tolerance of $thresh = 95\%$, which should be achieved at ($W_n = 14$).

  If we designate the believe states $S1$ and $S2$ respectively as the initial state (both Q-values and fl_τ resets, the network is assumed to have no stable flow during learning) and the flow harmony state, $S1$ is decided when the averaging function exceeds the threshold, which indicates that flow harmony has been achieved at least in the node's local interference group, otherwise the node resets to $S2$. In essence, every node has a window of 20 frames to isolate incoming reception slots and settle on a transmission slot. Whenever a particular node(s) fail(s) to settle and join the flow, the reset will make the node switch to another slot and potentially notify other nodes in the neighborhood as well.

$$fl_\tau \leftarrow (1-\gamma)fl_\tau + \lambda \tag{11}$$

  Where fl_τ,γ and λ denote the flow averaging, the learning/tracking rate and the increment scale respectively.

By using this two stage solution, Aloha-QUPAF unlike Aloha-Q effectively isolate both reception slots from transmission slots and find an implicit way of getting the feedback signal of node's action based on the individual nodes experiencing of successful reception of continuous stream of packets. Furthermore, it differs from framed-Aloha, since it can intelligently create and maintain a robust collision free schedule. The complete algorithm is given below.

Algorithm 1: Aloha-QUPAF Algorithm

```
Result: $S1$,or $S2$
S1;
Initialization;
$\alpha$, $\gamma$, $\lambda$, $\psi$ // From Table 5;
// For all n ;
$Q[n] \leftarrow rand([0,1))$;
$W_n \leftarrow 20$, $thresh \leftarrow 0.95$, $fl_\tau \leftarrow 0$;
S2;
while node is online do
    if Reception then
        get $rx\_s$ ;
        //Activating the packet reception flag;
        $Rx\_\tau \leftarrow True$;
        $Q[rx\_s] \leftarrow Q[rx\_s] + \alpha(\psi - Q[rx\_s])$;
    end
    // Frame Block;
    $W_n - = 1$;
    if $Rx\_\tau$ then
        $fl_\tau \leftarrow (1-\gamma)fl_\tau + \lambda$;
        $Q[tx\_s] \leftarrow Q[tx\_s] + \alpha(\psi - Q[tx\_s])$;
    end
    // Belief State Block: compares flow rate with threshold;
    if $W_n == 0$ then
        if $fl_\tau < thresh$ then
            // Node resets parameters;
            $node \leftarrow$ S1;
        else
            // Maintain parameters;
            $node \leftarrow$ S2;
        end
        $W_n = 20$;
    end
    // Transmission slot selection;
    $tx\_s \leftarrow [x | x \ni \text{argmax}_{x \in \mathcal{X}} Q[x]]$;
    //De-activating the packet reception flag;
    $Rx\_\tau \leftarrow False$;
end
```

Table 5: Simulation Parameters

Parameter	Value
Transmission/Reception Data Rate	640bps
Data Packet Size	632bits
ACK Packet Size	16bits
Slot Size	640bits
Slots per frame	4
Reception Range	200m
ψ	$\pm$ 1
α	0.1
λ	$\pm$ 0.1
γ	0.2
1 hop Propagation Delay (Relative to packet size)	1s

4.2. Results

Since the focus of this work is principally to improve performance in terms of channel utilization measured at the sink, Aloha-QUPAF is compared to a state-of-the-art Aloha-Q which employed a similar Q-learning technique, and a baseline framed-Aloha scheme in terms of the normalised utilization. We simulated networks of varying hop lengths with the protocols configured with respect to the structures in Figure 3. For fair comparison, as our proposed slot structure is constrained to $K\tau > 1$, we only compare Aloha-QUPAF with the other protocols in the $K\tau > 1$ regime. The network was simulated in the Riverbed Modeler (formerly OPNET) environment, the setup use the parameters given in Table 5 which are based on a modem developed by Newcastle University [37]. In all cases, the network is simulated for 15000 frames, with a single saturated source at one end of the network and a sink at the other end. In terms of result collection, due to the continuous nature of the learning of Aloha-QUPAF algorithm, results are collected from the beginning of the simulation.

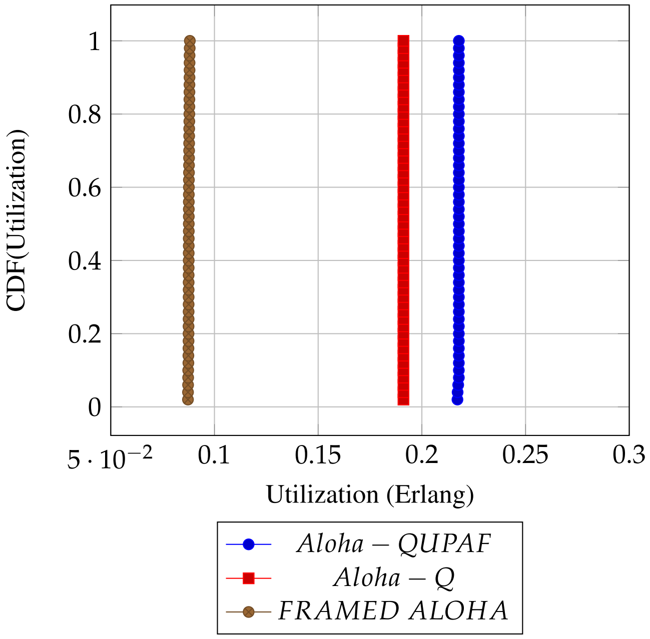


Figure 5. $K_\tau > 1$: 4 Hop utilisation performance comparison

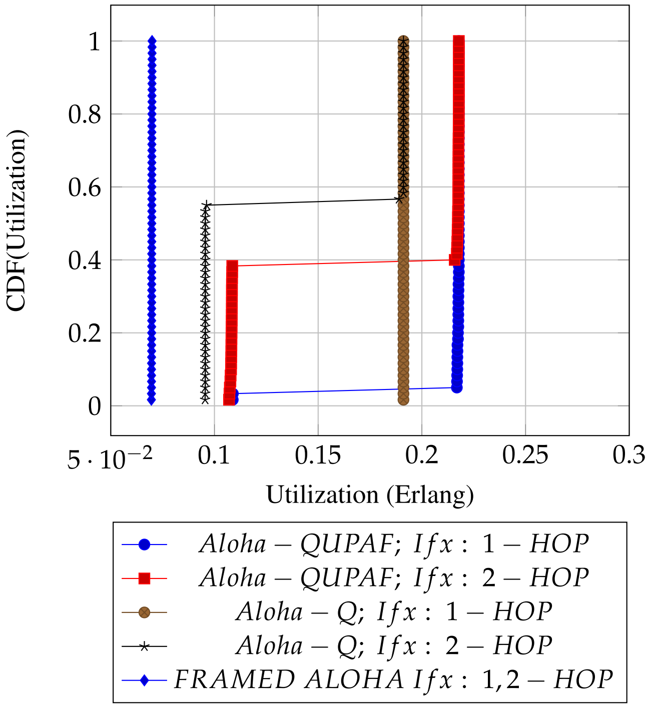


Figure 6. $K_\tau > 1$: 8 Hop utilisation performance comparison

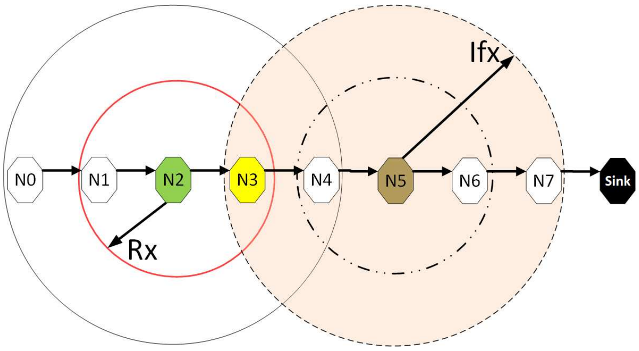


Figure 8. The hidden node problem

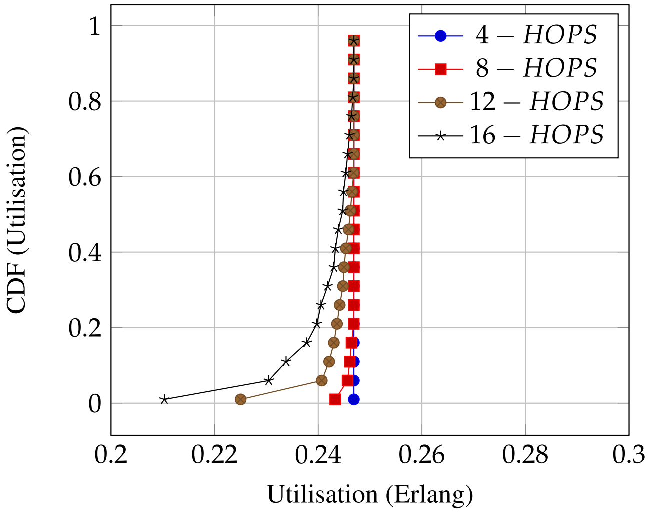


Figure 7. Aloha-QUPAF utilisation for 4,8,12, and 16 hops networks using the proposed slot structure

5. Discussion

Figure 5 and Figure 6 are results obtained when the network is simulated on 4 and 8 hop networks respectively. The figures compare the performance of Aloha-QUPAF with Aloha-Q and Framed Aloha. This comparison is particularly important as the protocols share similar reception conditions in the $K\tau > 1$ scenario; transmission and reception occur in the same slot (Figure 1). Evidently in this setup, both Aloha-QUPAF and Aloha-Q are dramatically affected as the network size increases (4-hops to 8-hops). The maximum utilisation of Aloha-QUPAF (0.217 Erlang) and Aloha-Q (0.191 Erlang) are both sharply halved for about 40% and 58% of the simulated cases respectively. This performance drop can be explained by the presence of the hidden-node phenomenon [38,39]. This is simply the situation whereby a particular communication between any two nodes is interfered by another transmission withing range of the receiver.

Figure 8 depicts the hidden-node problem in a 8-hop chain network, in a situation whereby both $N2$ and $N5$ share the same transmission slots, thus, transmission from $N2$ to $N3$ will be periodically interfered with by $N5$ transmitting to $N6$, as packets are relayed along the chain. The effect of the hidden-node problem as the reason for the performance degradation is confirmed by the agreement shown in the simulation results obtained when the interference range (*Ifx*) is reduced from 2-hops to 1-hop in the 8-hops chain (Figure 6) with the results in the 4-hops network (Figure 5). This is because, in a 2-hops interference range model, a 4 hop range chain network is of insufficient length for the issue to manifest. Mitigating the hidden node issue is a subject of further work. Another important metric worth mentioning is the end to end delay, however, it was not presented here, since Aloha-QUPAF does not implement packet retransmissions. Therefore, neglecting any processing and queuing delays in the nodes, the E2E delay is fixed as a function of the number of hops in the network. The simulations show that Aloha-QUPAF achieves 0.124

Erlangs at its worst and 0.248 Erlangs at its best, outperforming both Aloha-Q (0.19 Erlangs best) and framed Aloha (0.069 Erlangs) respectively by at least 13% and 148% in all simulated scenarios.

Figure 7 presents the performance of Aloha-QUPAF with our proposed slot structure (Figure 3) in the $K\tau < 1$ scenario. To demonstrate how Aloha-QUPAF protocol is affected by network length, we extend the range to 16 hops and evaluate its performance. The results show subtle drop in the overall performance from 4 to 16 hops. The decrease in performance is attributable to the increase in the hidden-node spots (bottlenecks points) and the time needed for the protocol to find collision free schedule as the network size increases. As each time a node switch to a different transmission slot, will have a ripple effect across the neighboring nodes, causing others to potentially switch slot as well. Essentially resetting the process. Despite a lack of explicit acknowledgement signal, the protocol demonstrates significant performance improvement with more than 90% of cases achieving 0.24 Erlang for networks in the 4-12 hops ranges, and 80% for the 16-hops range.

6. Conclusions

In this work, we present a simple slot structure based on the relationship between packet transmission duration and propagation delays in conjunction with two stage reinforcement learning techniques to develop a novel MAC protocol (Aloha-QUPAF) that can achieve near channel capacity utilisation in a UASN chain topology. Our solution addresses the excessive overhead required in slot structures used by typical slotted/framed protocols. Incorporating a Q-learning in the protocol makes it robust against network and channel changes due to the high dynamic underwater environment. Furthermore, one of the primary goals is for the protocol has to be distributed, adaptive, simple and computationally inexpensive so that it is suitable for use in cheap and low capacity modems.

To implement our solution, firstly, we analyse the slot structure using an intuitive diagrammatic representation to map the achievable channel utilisation levels. We then reformulate a Q-learning routine that exploit an implicit feedback signal to negatively reinforces and isolates reception slots in the slot selection phase. Secondly, by averaging the packet flow rate we are able to generate a distribution for belief states that control and consolidate the choice of transmission slot to achieve overall network wide packet flow. We finally evaluate and demonstrated that Aloha-QUPAF significantly outperforms the comparable protocols with similar Q-learning and slotting concepts.

Author Contributions: Conceptualization, I.A; methodology, I.A.; software, I.A.; validation, I.A. and P.M.; investigation, I.B.; writing—original draft preparation, I.B.; writing—review and editing, I.B. and P.M.; visualization, I.B.; supervision, P.M.; funding acquisition, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Nigerian Government through the Petroleum Technology Trust Fund (PTDF). The work of Professor Mitchell was in part supported by the U.K. Engineering and Physical Sciences Research Council through the USMART Project under Grant EP/P017975/1

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

458 ## Appendix A. Pictorial Analysis

Label	Description
▭	Packet
t	transmission
r	reception
i	interference
x	collision
***i**	ith packet
Fi	ith frame

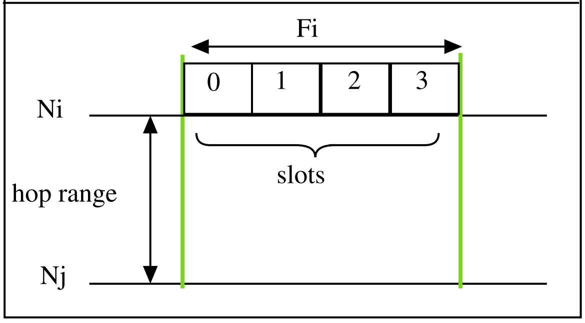


Figure A1. Legend for packet labels and illustrations

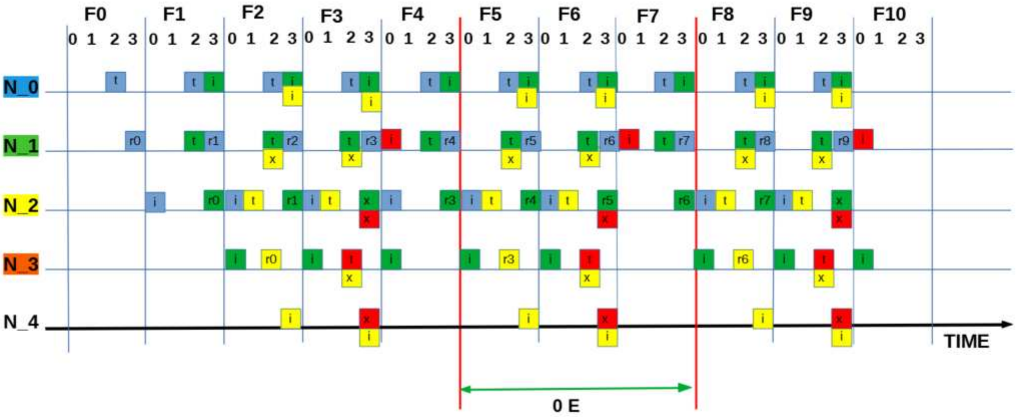


Figure A2. SEQUENCE:[2 2 1 2]: "Worst" measured utilisation based on zero packets get delivered = 0.0E

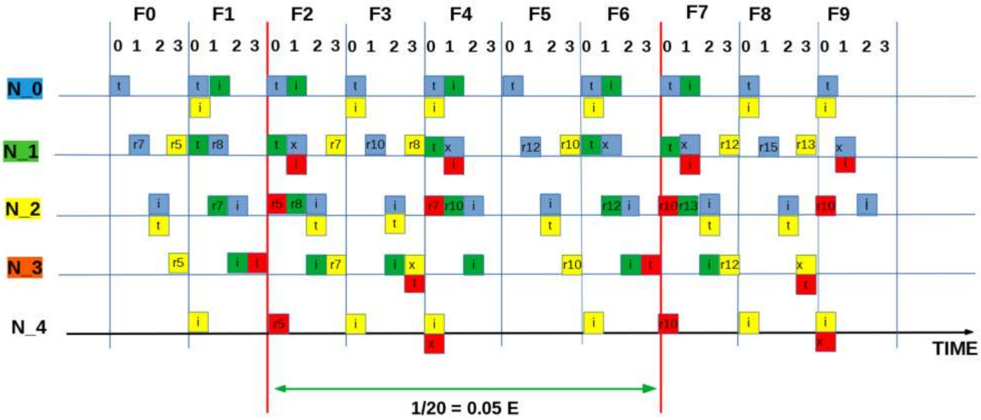


Figure A3. SEQUENCE:[0 0 2 3]: "Intermediate" measured utilisation based on one packet in five frames (20 slots) = 0.05E

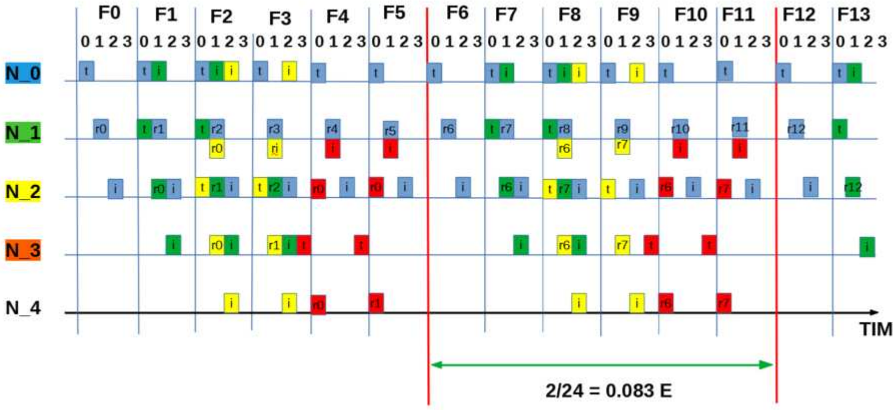


Figure A4. SEQUENCE:[0 0 0 3]: "Intermediate" measured utilisation based on two packets in six frames (24 slots) = 0.083E

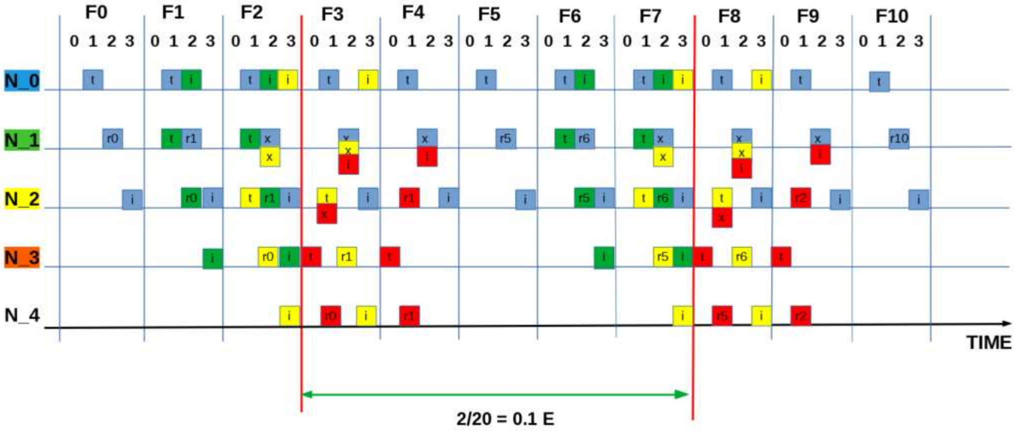


Figure A5. SEQUENCE:[1 1 1 0]: "Intermediate" measured utilisation based on two packets in five frames (20 slots) = 0.1E

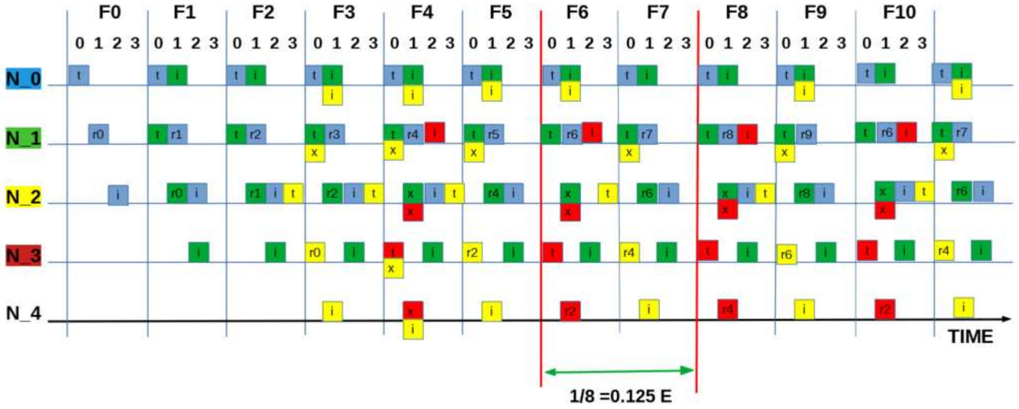


Figure A6. SEQUENCE:[0 0 3 0]: "Half" measured utilisation based on one packet every two frames (8 slots) = 0.125E

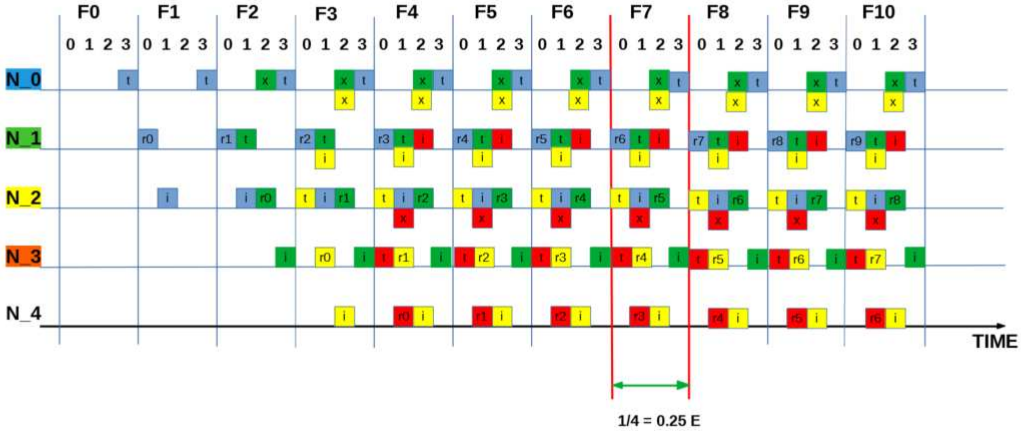


Figure A7. SEQUENCE:[0 2 1 1]: "Best" measured utilisation is one packet in every frames (4 slots) = 0.25E

References

1. Akyildiz, I.F.; Pompili, D.; Melodia, T. Underwater acoustic sensor networks: research challenges. *Ad Hoc Networks* **2005**, *3*, 257–279.
2. Syed, A.A.; Ye, W.; Heidemann, J. T-Lohi: A New Class of MAC Protocols for Underwater Acoustic Sensor Networks. 2008.
3. Noh, Y.; Shin, S.; Ieee. Survey on MAC Protocols in Underwater Acoustic Sensor Networks. *2014 14th International Symposium on Communications and Information Technologies (ISCIT)* **2014**, pp. 79–83.
4. Gkikopouli, A.; Nikolakopoulos, G.; Manesis, S. A survey on Underwater Wireless Sensor Networks and applications. 2012.
5. Otnes, R.; Asterjadhi, A.; Casari, P.; Goetz, M.; Husøy, T.; Nissen, I.; Rimstad, K.; van Walree, P.; Zorzi, M. *Underwater Acoustic Networking Techniques*; Springer, 2012.
6. Guan, J.; Huang, J.; Lu, J.; Wang, J. The underlying design in underwater acoustic wireless sensor network. 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), 2013, pp. 1–5.
7. Yang, X. Underwater Acoustic Sensor Networks **2010**.
8. Stoianov, I.; Nachman, L.; Madden, S.; others. PIPENET: A wireless sensor network for pipeline monitoring. *Int. J. Distrib. Sens. Netw.* **2007**.
9. Gao, M.S.; Li, J.; Li, W.; Deng, Z.X.; Ieee. *A Multi-Channel MAC Protocol for Underwater Acoustic Networks*; 2015 Ieee 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks, Ieee: New York, 2015.
10. Chen, K.; Ma, M.; Cheng, E.; Yuan, F.; Su, W. A Survey on MAC Protocols for Underwater Wireless Sensor Networks. *IEEE Communications Surveys & Tutorials* **2014**, *16*, 1433–1447.
11. Karl, H.; Willig, A. *Protocols and Architectures for Wireless Sensor Networks*; Wiley, 2005.
12. Morozs, N.; Mitchell, P.; Zakharov, Y.V. TDA-MAC: TDMA Without Clock Synchronization in Underwater Acoustic Networks. *IEEE Access* **2018**, *6*, 1091–1108.
13. Morozs, N.; Mitchell, P.D.; Zakharov, Y. Linear TDA-MAC: Unsynchronized Scheduling in Linear Underwater Acoustic Sensor Networks, 2019.
14. Abramson, N. Development of the ALOHANET. *IEEE Transactions on Information Theory* **1985**, *31*, 119–123. doi:10.1109/TIT.1985.1057021.

15. Zhou, Y.; Chen, K.; He, J.; Guan, H. Enhanced Slotted Aloha Protocols for Underwater Sensor Networks with Large Propagation Delay. 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), 2011, pp. 1–5.
16. Chirdchoo, N.; Soh, W.S.; Chua, K.C. Aloha-Based MAC Protocols with Collision Avoidance for Underwater Acoustic Networks. IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications, 2007, pp. 2271–2275.
17. Chirdchoo, N.; s. Soh, W.; Chua, K.C. RIPT: A Receiver-Initiated Reservation-Based Protocol for Underwater Acoustic Networks. *IEEE J. Sel. Areas Commun.* **2008**, *26*, 1744–1753.
18. Gorma, W.; Mitchell, P.D.; Morozs, N.; Zakharov, Y.V. CFDAMA-SRR: A MAC Protocol for Underwater Acoustic Sensor Networks. *IEEE Access* **2019**, *7*, 60721–60735.
19. Kredo, K.B.; Mohapatra, P. A hybrid medium access control protocol for underwater wireless networks. Proceedings of the second workshop on Underwater networks; Association for Computing Machinery: New York, NY, USA, 2007; WuWNet '07, pp. 33–40.
20. Yan, Y.; Mitchell, P.; Clarke, T.; Grace, D. Adaptation of the ALOHA-Q protocol to Multi-Hop Wireless Sensor Networks. European Wireless 2014; 20th European Wireless Conference, 2014, pp. 1–6.
21. Alhassan, I.; Mitchell, P. Monitoring free span sections of subsea pipeline with ALOHA-Q. 2019. URSI Festival of Radio Science ; Conference date: 16-12-2019 Through 16-12-2019.
22. Chu, Y.; Kosunalp, S.; Mitchell, P.; Grace, D.; Clarke, T. Application of reinforcement learning to medium access control for wireless sensor networks. *Engineering Applications of Artificial Intelligence* **2015**, *46*, 23–32. doi:10.1016/j.engappai.2015.08.004.
23. Mickus, T.; Mitchell, P.; Clarke, T. The emergence MAC (E-MAC) protocol for wireless sensor networks. *Engineering Applications of Artificial Intelligence* **2017**, *62*, 17–25. © 2017 Elsevier Ltd. This is an author-produced version of the published paper. Uploaded in accordance with the publisher's self-archiving policy., doi:10.1016/j.engappai.2017.03.003.
24. Wang, Y.; He, H.; Tan, X. Robust Reinforcement Learning in POMDPs with Incomplete and Noisy Observations **2019**. [arXiv:cs.LG/1902.05795].
25. Rappaport, T. Wireless communications principles and practice edition **2001**.
26. Mandal, P.; De, S.; Chakraborty, S.S. Characterization of Aloha in underwater wireless networks. 2010 National Conference On Communications (NCC), 2010, pp. 1–5.
27. Lucani, D.E.; Stojanovic, M.; Medard, M. On the Relationship between Transmission Power and Capacity of an Underwater Acoustic Communication Channel. OCEANS 2008 - MTS/IEEE Kobe Techno-Ocean, 2008, pp. 1–6.
28. Sen, S.; Dorsey, D.J.; Guérin, R.; Chiang, M. Analysis of Slotted Aloha with multipacket messages in clustered surveillance networks. MILCOM 2012 - 2012 IEEE Military Communications Conference, 2012, pp. 1–6.
29. Ma, R.T.B.; Misra, V.; Rubenstein, D. An Analysis of Generalized Slotted-Aloha Protocols. *IEEE/ACM Trans. Netw.* **2009**, *17*, 936–949.
30. Sutton, R.S.; Barto, A.G. Reinforcement learning: An introduction **1998**.
31. Claus, C.; Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* **1998**, *1998*, 2.
32. Boutilier, C. Sequential optimality and coordination in multiagent systems. IJCAI, 1999, Vol. 99, pp. 478–485.
33. Robbins, H.; Monro, S. A stochastic approximation method. In *Herbert Robbins Selected Papers*; Springer, 1985; pp. 102–109.
34. Oliehoek, F.A.; Amato, C. *A Concise Introduction to Decentralized POMDPs*; Springer, Cham, 2016.
35. Wiering, M.; van Otterlo, M. *Reinforcement Learning: State-of-the-Art*; Springer Science & Business Media, 2012.
36. Bonabeau, E.; Dorigo, M.; Theraulaz, G. *Swarm Intelligence: From Natural to Artificial Systems*; OUP USA, 1999.
37. USMART - Newcastle University.
38. Tanenbaum. *Computer Networks*; Pearson Education(singapore) Pte. Limited, 2011.
39. Peterson, L.L. *Computer Networks - A Systems Approach 3rd Edition*.