Multi-Armed-Bandit-Based Dynamic Optimization for Coexisting IEEE 802.11ah and IEEE 802.15.4g Networks in Sub-GHz Band

Hiroya Kai[1], Song-Ju Kim[2,1], Maki Arai[3], Jin Nakazato[1], Mikio Hasegawa[1]

[1]*Department of Electrical Engineering, Tokyo University of Science,* Tokyo, Japan

[2]*SOBIN Institute LLC,* Hyogo, Japan

[3]*Department of Engineering, Shibaura Institute of Technology,* Tokyo, Japan

***Abstract*—** Coexistence of IEEE 802.11ah (Wi-Fi HaLow) and IEEE 802.15.4g (Wi-SUN) in the sub-GHz band leads to cross-technology interference that degrades packet delivery rate (PDR) and fairness, especially under dense, duty-cycle-limited deployments. Existing coexistence recommendations mainly provide static parameter settings and do not enable dynamic, device-level adaptation. To address this gap, we propose a machine-learning-based framework that formulates the joint channel and packet-size selection at each end node (EN) as a multi-armed bandit (MAB) problem and optimizes per-device PDR in a mixed 920 MHz deployment. Using QualNet 9.0 models of coexisting IEEE 802.11ah and IEEE 802.15.4g networks, we implement and compare three online learning policies—ε-greedy, UCB-1 tuned with exponential forgetting, and a congestion-aware continuous-value Tug-of-War (ToW)—with normalized PDR as the reward in 10-minute learning rounds. Simulation results show that the proposed ToW policy converges faster than ε-greedy and UCB-1 tuned, maintains comparable or higher average PDR, and significantly improves the PDR of previously low-performing IEEE 802.15.4g nodes. These findings demonstrate that per-device MAB-based learning is an effective mechanism for dynamic coexistence management and provide design insights for applying MAB-based adaptation in heterogeneous LPWA deployments.

Keywords—Low-Power-Wide-Area Networks, Massive IoT, Multi-Armed Bandit, Sub-GHz Coexistence.

I. Introduction

Driven by the rapid proliferation of Internet-of-Things (IoT) devices, the number of connected endpoints reached 16.6 billion by the end of 2023, of which Low-Power Wide-Area Network (LPWAN) links accounted for roughly 8% (≈1.3 billion)[1]. By 2030, total IoT connections are projected to grow to about 40 billion, with the LPWA share rising to around 10% [1]. In parallel, the LPWA market is expected to expand from USD 6.5 billion in 2023 to USD 48.1 billion by 2030 (2024–2030 CAGR ≈ 33.1%) [2]. This diffusion has led to the coexistence of multiple LPWA protocols such as Wi-SUN (IEEE 802.15.4g), Wi-Fi HaLow (IEEE 802.11ah), LoRaWAN, Sigfox, NB-IoT, and LTE-M. Notably, LoRa, IEEE 802.11ah, and IEEE 802.15.4g commonly share sub-GHz ISM bands (Japan: 920 MHz; EU: 868 MHz; North America: 915 MHz; parts of Asia: 433 MHz), forming a shared-spectrum ecosystem [3]. The growing protocol diversity and device density intensify channel contention and cross-technology interference, elevating the risk of degraded throughput, PDR, and latency.

In sub-GHz coexistence between IEEE 802.11ah and IEEE 802.15.4g, severe interference from 802.11ah to IEEE 802.15.4g has been reported even within standardized coexistence mechanisms. In response, the IEEE 802.19.3 working group has conducted extensive coexistence studies and published recommendations that specify suitable network profiles in terms of network size, offered load, frame size, and MAC backoff parameters for different deployment scenarios [5]. However, these network-profile-based recommendations are essentially static: once the load and size category of a deployment are chosen, the corresponding frame-size and backoff settings remain fixed and are not designed to track fast, device-level fluctuations of interference and traffic load.

To address these limitations, we propose a per-device online learning algorithm that dynamically selects the uplink channel and packet size of each end node (EN) based on its locally observed PDR. The algorithm is formulated within a multi-armed bandit (MAB) framework and aims to maximize the long-term PDR of each EN while alleviating the unfair impact of IEEE 802.11ah on IEEE 802.15.4g under coexistence. We implement and evaluate this approach in QualNet [6] by modeling a mixed deployment where IEEE 802.11ah and IEEE 802.15.4g share spectrum and compare several learning-based decision policies.

The remainder of this paper is organized as follows. Section II introduces the system model and formulates the per-device channel and packet-size optimization problem under IEEE 802.11ah / IEEE 802.15.4g coexistence. Section III presents the proposed MAB-based learning policies and their implementation. Section IV describes the simulation setup and reports experimental results. Section V concludes the paper and outlines future research directions.

II. System Model and Problem Formulation

A. Network Topology and Traffic Model

We consider a mixed sub-GHz deployment where IEEE 802.11ah (Wi-Fi HaLow) and IEEE 802.15.4g (Wi-SUN) share the same 920 MHz band. End nodes (ENs) of both technologies periodically transmit uplink traffic to their associated access points (APs) using carrier-sense-based

channel access under duty-cycle-limited operation, which is representative of dense industrial IoT and utility-network deployments. Each EN selects a channel in the 920 MHz band and a packet size from a given set of candidates.

In this coexistence setting, PHY/MAC-level asymmetries between IEEE 802.11ah and IEEE 802.15.4g inherently bias channel access. IEEE 802.11ah typically employs a higher energy-detect (ED) threshold than IEEE 802.15.4g, so an IEEE 802.15.4g frame that is decodable at an IEEE 802.15.4g receiver may still be invisible to IEEE 802.11ah ED-CCA; an 802.11ah node can therefore start a new transmission and destroy the ongoing IEEE 802.15.4g frame. Moreover, although both standards use carrier sense with backoff, the effective backoff progression of 802.11ah is often faster, allowing 802.11ah to preempt the medium. In addition, IEEE 802.15.4g usually operates at lower PHY rates than IEEE 802.11ah, so each IEEE 802.15.4g frame occupies the channel for a longer time and increases contention under mixed traffic.

As a result, IEEE 802.15.4g nodes tend to experience degraded PDR and unfair access compared with IEEE 802.11ah nodes, especially under moderate-to-high duty cycles. This work aims to mitigate these coexistence problems by dynamically optimizing the channel and packet-size selection at each EN under a fixed network-level duty-cycle constraint.

B. Problem Formulation

In the coexistence scenario described above, each EN must choose a channel–packet-size configuration for its uplink transmissions in the sub-GHz band. Under a fixed network-level duty-cycle constraint, inappropriate joint configurations across ENs can severely degrade the PDR of IEEE 802.15.4g nodes and aggravate unfairness between the two technologies. Our goal is to design a distributed online adaptation mechanism that allows each EN to adjust its own configuration based only on locally observable performance. We assume an autonomously distributed network in which there is no central controller and each EN has no information about the existence, actions, or PDR of other ENs, including those belonging to the other technology.

We model this per-device adaptation as a sequence of discrete decision epochs $t = 1,2,\dots,T$. At the beginning of epoch t, EN j selects a configuration $x_j(t) = (f_j(t), s_j(t))$ from its finite action set A_j, which collects all feasible channel–packet-size combinations for that EN. During epoch t, EN j uses $x_j(t)$ for its uplink transmissions. At the end of the epoch, EN j measures its packet delivery rate $\mathrm{PDR}_j(t)$ and defines a normalized reward $r_j(t) = \frac{\mathrm{PDR}_j(t)}{100} \in [0,1]$, which we use as the sole performance indicator.

The design objective is to construct a distributed online adaptation rule for each EN that maps its own past observations $\{x_j(\tau), r_j(\tau)\}_{\tau=1}^{t}$ to the next configuration $x_j(t+1)$, so as to maximize its long-term normalized PDR. Equivalently, each EN seeks to maximize the cumulative reward $\sum_{t=1}^{T} r_j(t)$ (or time-averaged reward $\frac{1}{T}\sum_{t=1}^{T} r_j(t)$ while implicitly mitigating coexistence-induced unfairness between IEEE 802.11ah and IEEE 802.15.4g nodes.

We model the above per-device adaptation problem within a multi-armed bandit framework, where each feasible configuration (channel, packet size) is treated as one arm and the normalized per-round PDR $r_j(t)$ is used as the reward. In Section III, we instantiate this formulation and compare ε-greedy and UCB-1 tuned with forgetting, together with a continuous-value Tug-of-War (ToW) based policy tailored to the joint channel–packet-size selection problem under IEEE 802.11ah / IEEE 802.15.4g coexistence.

III. Proposed MAB-Based Learning Policies

A. Common Per-Device Bandit Framework

In the per-device optimization problem of Section II, each EN chooses a configuration consisting of a channel and a packet size for its uplink transmissions. We discretize this configuration space into a finite set of candidate actions and treat each action as one arm in a bandit model. For each EN j, let K_j denote the number of available arms, and index the arms by an integer k in $\{1, \dots, K_j\}$. The mapping between an arm index k and the corresponding (channel, packet-size) pair is fixed for each EN and is shared across all learning policies.

For each EN j, the learning policy maintains per-arm statistics that are updated once per learning round. At the end of round t, for every arm k, we store the cumulative reward $G_{j,k}(t)$, defined as the sum of all rewards obtained when arm k has been selected by EN j up to round t, and an effective pull count $N_{j,k}(t)$. The empirical mean reward of arm k for EN j is then given by

$$\hat{\mu}_{j,k}(t) = \begin{cases} \dfrac{G_{j,k}(t)}{N_{j,k}(t)}, & N_{j,k}(t) > 0 \\ 0, & N_{j,k}(t) = 0 \end{cases} \quad (1)$$

At the beginning of each round t, EN j selects its next arm $a_j(t)$ by applying its policy to its own per-arm statistics. At the end of the round, EN j observes the reward $r_j(t)$ and updates the statistics of the selected arm. The three policies considered in this paper—ε-greedy, UCB-1 tuned with exponential forgetting, and continuous-value ToW—share this common structure and differ only in how they define the selection rule and, in the case of UCB-1 tuned and ToW, how they modify the statistics.

B. ε-Greedy Policy

The ε-greedy policy balances exploitation of the currently best arm and exploration of other arms. When EN j selects arm $a_j(t)$ and observes reward $r_j(t)$, it updates the cumulative reward $G_{j,a_j(t)}$ and the pull count $N_{j,a_j(t)}$ of the selected arm as

$$G_{j,a_j(t)}(t+1) = G_{j,a_j(t)}(t) + r_j(t) \quad (2)$$

$$N_{j,a_j(t)}(t+1) = N_{j,a_j(t)}(t) + 1 \quad (3)$$

and the empirical mean reward $\hat{\mu}_{j,k}(t)$ is computed as in (1).

At the beginning of round $t+1$, EN j draws a uniform random number $u \sim \mathcal{U}(0,1)$. With probability ε, EN j

explores by choosing a random arm; otherwise, it exploits the arm with the largest empirical mean reward:

$$a_j(t+1) = \begin{cases} \text{random arm in} \{1, \dots, K_j\}, & \text{with probability } \varepsilon \\ \arg\max_k \hat{\mu}_{j,k}(t), & \text{with probability } 1-\epsilon \end{cases} \quad (4)$$

C. UCB-1 Tuned Policy

UCB-1 tuned selects arms based on an optimism-in-the-face-of-uncertainty index, with an exploration bonus that depends on an upper confidence bound of the variance. To cope with non-stationary interference conditions in the coexistence scenario, we incorporate exponential forgetting into both the cumulative rewards and the pull counts. Let $\alpha \in (0,1]$ denote the forgetting factor. When EN j selects arm $a_j(t)$ and observes reward $r_j(t)$, we first apply forgetting to all arms:

$$G_{j,k(t)}(t^+) = \alpha G_{j,k(t)}(t), N_{j,k(t)}(t^+) = \alpha N_{j,k(t)}(t), \forall k \quad (5)$$

and then update the chosen arm as

$$G_{j,k(t)}(t+1) = G_{j,k(t)}(t^+) + r_j(t) \quad (6)$$

$$N_{j,k(t)}(t+1) = N_{j,k(t)}(t^+) + 1 \quad (7)$$

The empirical mean reward is

$$\hat{\mu}_{j,k}(t+1) = \frac{G_{j,k}(t+1)}{N_{j,k}(t+1)} \quad (8)$$

During the initial exploration phase, any arm with $N_{j,k}(t+1) = 0$ is prioritized and selected at least once. After all arms have been tried, we define the total effective number of pulls as

$$N_j^{tot}(t+1) = \sum_{k=1}^{K_j} N_{j,k}(t+1) \quad (9)$$

Following the UCB-1 tuned formulation, we approximate the variance proxy of arm k as.

$$v_{j,k}(t+1) = \left(\frac{G_{j,k}(t+1)}{N_{j,k}(t+1)}\right)^2 - \hat{\mu}_{j,k}^2(t+1) + \sqrt{\frac{2logN_j^{tot}(t+1)}{N_{j,k}(t+1)}} \quad (10)$$

and define the exploration bonus as

$$B_{j,k}(t+1) = \sqrt{\frac{logN_j^{tot}(t+1)}{N_{j,k}(t+1)} \cdot \min\{0.25, v_{j,k}(t+1)\}} \quad (11)$$

The UCB-1 tuned index is then

$$I_{j,k}(t+1) = \hat{\mu}_{j,k}(t+1) + B_{j,k}(t+1) \quad (12)$$

and the next arm is chosen as

$$a_j(t+1) = \arg\max_k I_{j,k}(t+1) \quad (13)$$

D. Continuous Tug-of-War Policy

ToW dynamics were originally proposed by Kim as a bio-inspired framework for distributed reinforcement learning [9]. Building on this concept, both the binary and real-valued reward versions of ToW [10, 11] were introduced, and it was demonstrated that these methods can achieve efficient channel selection with very low computational complexity [12]. Refs [13] and [14] extended this idea to distributed channel selection in massive IEEE 802.15.4g IoT networks by incorporating forgetting factors to cope with dense and time-varying interference. Urabe et al. [15] applied autonomous distributed reinforcement learning with ToW dynamics to spreading-factor selection in LoRa networks and confirmed that ToW-based policies can adapt SF to distance and SNR without centralized coordination.

More recently, a continuous-value variant of ToW (CToW) was proposed to enable real-valued updates of arm preferences in dynamic radio environments, improving adaptability over conventional discrete-value ToW implementations. CToW enhances stability and reactivity by computing scores as a combination of cumulative reward and congestion-dependent penalties, and introduces Gaussian noise to promote exploration [16]. A formal regret or convergence analysis of the CToW-based policy is left as future work.

Motivated by these studies, we design a continuous-value ToW policy tailored to our per-device bandit framework, where each arm corresponds to a joint choice of uplink channel and packet size in the mixed IEEE 802.11ah / IEEE 802.15.4g deployment. For each EN j and arm k, we reuse the cumulative reward $G_{j,k}(t)$ defined in Section III-A.

When EN j selects arm $a_j(t)$ and observes reward $r_j(t)$, the counters are updated as

$$G_{j,a_j(t)}(t+1) = G_{j,a_j(t)}(t) + r_{j(t)} \quad (14)$$

$$N_{j,a_{j(t)}}(t+1) = N_{j,a_{j(t)}}(t) + 1 \quad (15)$$

The empirical gain of each arm is

$$P_{j,k}(t+1) = \begin{cases} \frac{G_{j,k}(t+1)}{N_{j,k}(t+1)}, & N_{j,k}(t+1) > 0 \\ 0, & N_{j,k}(t+1) = 0. \end{cases} \quad (16)$$

Let $P_{j,(1)}(t+1)$ and $P_{j,(2)}(t+1)$ denote the largest and second-largest elements of $\{P_{j,k}(t+1)\}_k$, respectively. We define

$$I_{j,k}(t+1) = \hat{\mu}_{j,k}(t+1) + B_{j,k}(t+1) \quad (17)$$

which acts as a congestion coefficient. The base ToW score of arm k is

$$q_{j,k}(t+1) = G_{j,k}(t+1) - \frac{\gamma_j(t+1)}{2} N_{j,k}(t+1) \quad (18)$$

The first term accumulates gains, whereas the second term penalizes heavily used arms with a coefficient proportional to the combined expected gains of the two currently most promising arms. To enhance comparability across arms, we mean-center the scores and add Gaussian exploration noise:

$$X_{j,k}(t+1) = \left(q_{j,k}(t+1) - \bar{q}_j(t+1)\right) + \sigma\xi_{j,k}(t+1) \quad (19)$$

where $\bar{q}_j(t+1)$ is the average of $\{q_{j,k}(t+1)\}_k$, $\xi_{j,k}(t+1) \sim \mathcal{N}(0,1)$, and σ is the exploration-noise amplitude. The next arm is chosen as

$$a_{j(t+1)} = \arg\max_k X_{j,k}(t+1) \quad (20)$$

IV. Experiments and Results

A. Simulation and Learning Setup

We use the QualNet 9.0 network simulator[6] and implement PHY/MAC models that emulate IEEE 802.11ah and IEEE 802.15.4g according to the parameters in Table I. A single access point (AP) is placed within 10 m of the center of the simulation area. For each technology, 15 ENs are randomly and uniformly distributed in an annulus with inner radius 100 m and outer radius 500 m around the AP, as illustrated in Fig. 1.

For each experiment, the overall duty cycle of each network is fixed to 10% and applied uniformly to all ENs of that technology by setting the offered load per EN such that their aggregate duty cycle equals 10%. One learning round corresponds to a 10-minute simulation interval in QualNet. At the end of each round t, each EN j measures its PDR and converts it into the normalized reward

$$r_j(t) = \frac{\mathrm{PDR}_j(t)}{100} \in [0,1] \qquad (21)$$

which we use as the sole performance indicator. Unless otherwise noted, each learning policy is trained for $T = 1000$ rounds. This single-AP, 10%-duty-cycle configuration already yields severe coexistence-induced degradation for IEEE 802.15.4g nodes, and is therefore used as a baseline to isolate the impact of the learning policies; extensions to other duty cycles, traffic loads, node densities, and multi-AP topologies are left for future work.

The action spaces follow this configuration. Each IEEE 802.11ah EN has 30 arms corresponding to five payload sizes {200, 400, 600, 800, 1000} bytes combined with six 1 MHz channels in the 920 MHz band. Each IEEE 802.15.4g EN has 70 arms obtained from the same five payload sizes and 14 channels in the 920 MHz band.

For the UCB-1 tuned policy, we introduce an initial exploration phase: before applying the UCB-1 tuned selection rule, each arm of each EN is forced to be selected exactly five times so that all arms start with the same number of observations. After this phase, the policy switches to the UCB-1 tuned rule described in Section III-C. The hyperparameters of the three learning policies are fixed as follows. For the ε-greedy policy, ε is set to 0.2. For UCB-1 tuned, we use a forgetting factor $\alpha = 0.99$ together with the above exploration scheme. For the ToW policy, the exploration-noise amplitude σ is set to 0.001, which we empirically found to balance exploitation and exploration.

Table I Simulation Parameters

Protocol	**802.11ah**	**802.15.4g**
Frequency Bandwidth	1 MHz	281 kHz
Throughput	300 kbps	100 kbps
Tx Power	13 dBm	13 dBm
Rx Threshold	-95 dBm	-93 dBm

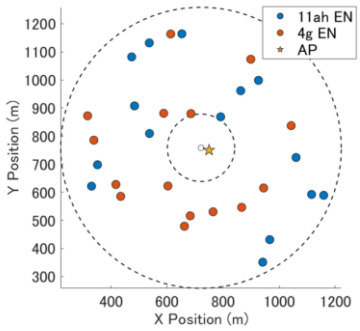

Fig.1: Node Placement

B. PDR Learning Curves and Convergence Speed

Figs. 2 and 3 show the evolution of the average per-EN PDR for the three policies under the 10% network duty-cycle scenario. PDR curves are smoothed using a moving average with a window of 20 rounds. In each figure, the solid lines show the smoothed PDR, while the faint dotted lines in the background indicate the instantaneous per-round PDR without smoothing. All reported mean values are computed from the unsmoothed samples. For IEEE 802.11ah (Fig. 2), the proposed ToW policy rapidly converges to almost 100% PDR within the first several tens of rounds and then remains essentially flat. UCB-1 tuned also approaches a PDR close to 100% in the long run, but its trajectory contains several pronounced drops in the middle rounds due to aggressive exploration. In contrast, ε-greedy only gradually improves from about 60% to the low-80% range and never reaches the high PDR region achieved by ToW and UCB-1 tuned.

For IEEE 802.15.4g (Fig. 3), ToW again converges quickly and stably, attaining a PDR close to 100%. ε-greedy converges much more slowly and saturates around 80% PDR. UCB-1 tuned suffers from an extended period of very low PDR in the middle of learning before recovering to around 90%. Consequently, its final PDR for IEEE 802.15.4g remains clearly lower than that of ToW, indicating that the proposed policy provides the best overall reliability, especially for the more vulnerable IEEE 802.15.4g ENs.

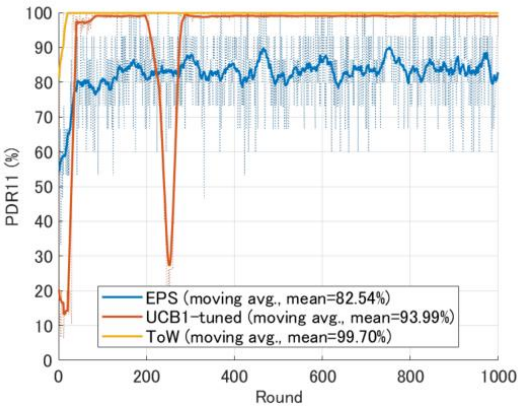

Fig.2: PDR learning curves for 802.11ah

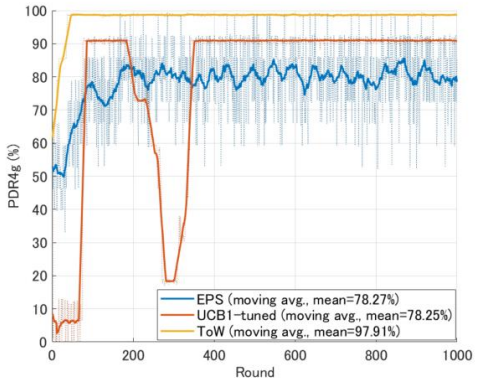

Fig.3: PDR learning curves for 802.15.4g

C. Channel and Packet-Size Selection Patterns

Figs. 4–7 show heatmaps of channel and packet-size choices over the learning rounds. In each figure, the horizontal axis is the round index $t = 1, \ldots, 1000$, and the vertical axis has 15 rows, each corresponding to one EN of the considered technology (IEEE 802.11ah in Figs. 4 and 5, IEEE 802.15.4g in Figs. 6 and 7). The three panels from top to bottom represent ε-greedy (EPS), UCB-1 tuned, and the proposed ToW policy. In the channel-selection figures (Figs. 4 and 6), the color denotes the selected center frequency (MHz), whereas in the packet-size figures (Figs. 5 and 7) it denotes the selected payload size (bytes).

With ε -greedy, the heatmaps remain highly mixed, indicating that both technologies keep switching among many channels and packet sizes and do not clearly converge. UCB-1 tuned produces horizontal stripes in the channel plots, showing that most ENs eventually settle on quasi-fixed channels, but the packet-size patterns remain scattered. In contrast, the proposed ToW policy yields stable horizontal bands in both channel and packet-size figures, meaning that ENs converge to consistent channel–packet-size configurations, which matches the PDR and fairness improvements in Figs. 2 and 3.

D. Per-Node Fairness and Spatial PDR Distribution

Figs. 8-10 show the spatial distribution of the average PDR per transmitter over the last 200 learning rounds at a 10% network duty cycle for ε-greedy, UCB-1 tuned, and ToW, respectively. Circles denote IEEE 802.11ah ENs and triangles denote IEEE 802.15.4g ENs, and the color indicates the average PDR. With ε-greedy (Fig. 8), most ENs achieve only moderate PDR values of about 70–85%, and there is noticeable variation among nodes. Several IEEE 802.15.4g ENs located near the outer ring exhibit lower PDR than the others. Under UCB-1 tuned (Fig. 9), many ENs reach high PDR values around 90–95%. However, one IEEE 802.15.4g EN suffers from almost zero PDR, which leads to severe unfairness despite the high network average. With the proposed ToW policy (Fig. 10), all ENs achieve uniformly high PDR, typically above 95%, and the difference between the best and worst nodes becomes very small. In particular, the IEEE 802.15.4g ENs that had low or even zero PDR under the baseline policies now attain PDR close to those of the IEEE 802.11ah ENs. This demonstrates that the proposed ToW dynamics not only improve the overall reliability but also enhance per-node fairness in the mixed IEEE 802.11ah / IEEE 802.15.4g deployment. In ToW, the volume-conservation term effectively penalizes arms that are heavily used across the network, so each EN is gradually pushed away from congested channels and packet sizes. As a result, the per-EN PDR values concentrate at similarly high levels (Fig. 10), whereas under ε-greedy and UCB-1 tuned a subset of nodes remains trapped in persistently low-PDR configurations (Figs. 8 and 9).

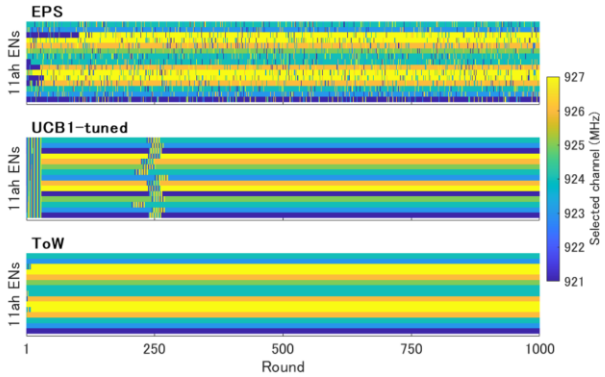

Fig.4: Channel selection patterns of IEEE 802.11ah ENs

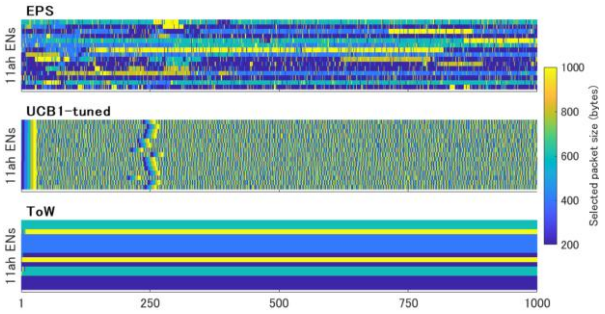

Fig.5: Packet-size selection patterns of IEEE 802.11ah ENs

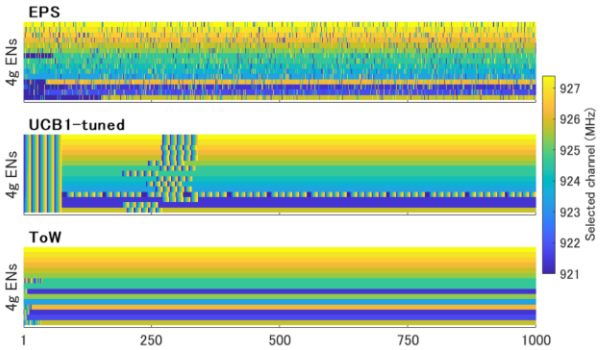

Fig.6: Channel selection patterns of IEEE 802.15.4g ENs

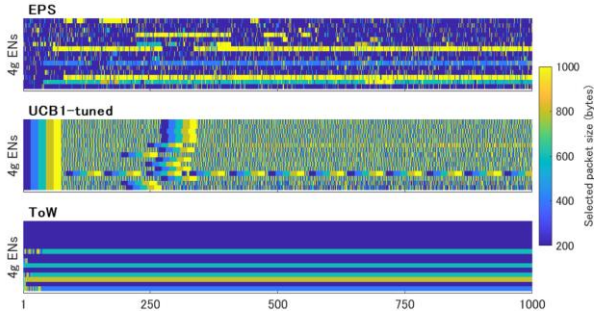

Fig.7: Packet-size selection patterns of IEEE 802.15.4g ENs

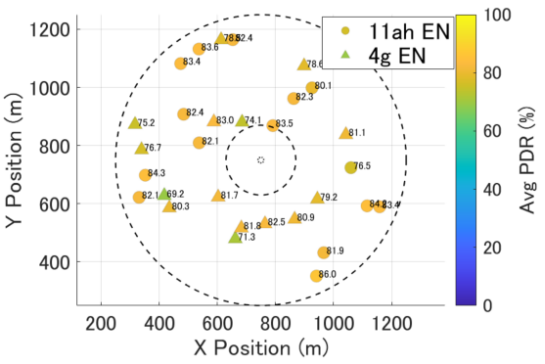

Fig.8: Average PDR by Transmitter and Node Layout (ε-greedy)

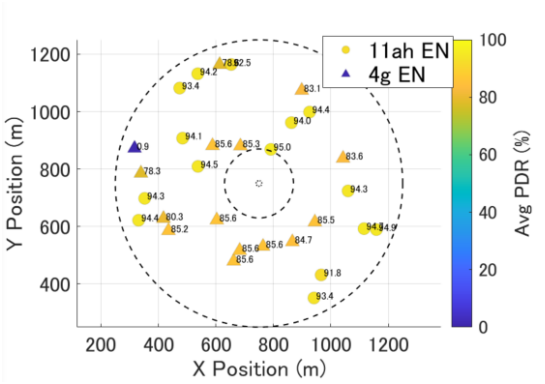

Fig.9: Average PDR by Transmitter and Node Layout (UCB-1 tuned)

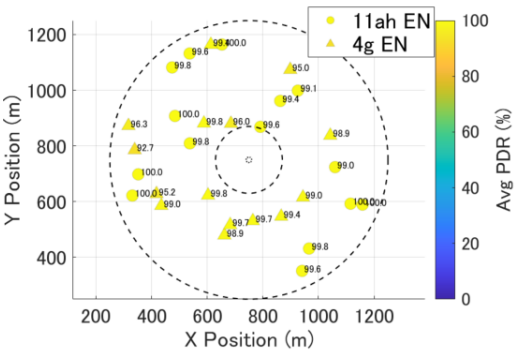

Fig.10: Average PDR by Transmitter and Node Layout (ToW)

V. Conclusion

In this paper, we formulated per-device channel and packet-size selection as an online learning problem and compared three MAB-based policies, including a proposed continuous ToW scheme tailored to IEEE 802.11ah / IEEE 802.15.4g coexistence in the sub-GHz band. Using a QualNet-based model of a single-AP topology at a 10% network duty cycle, the proposed ToW policy achieved the highest PDR for both technologies. For IEEE 802.11ah ENs, ToW and UCB-1 tuned eventually converged to PDR values close to 100%, while ε-greedy saturated at a lower level. For IEEE 802.15.4g ENs, ToW clearly outperformed the baselines, reaching near-100% PDR where ε-greedy and UCB-1 tuned remained significantly lower. The spatial analysis further showed that ToW improved per-node fairness, eliminating the severely degraded IEEE 802.15.4g nodes observed under the baseline policies.

Future work includes evaluating the proposed learning framework under more dynamic environments, such as time-varying traffic loads, interference patterns, and topology changes; assessing its performance under different node densities and spatial configurations; and generalizing the learning-based parameter adaptation to other sub-GHz coexistence scenarios involving protocols such as LoRa and ZigBee.

Acknowledgment

This work was supported partly by JSPS KAKENHI Grant Number JP22H05197.

References

[1] IoT Analytics, “Global IoT market forecast (in billions of connected IoT devices),”[Online]. Available: https://iot-analytics.com/number-connected-iot-devices/

[2] Next Move Strategy Consulting, “LPWA (Low-Power-Wide-Area) Market,”2023–2030forecast.” [Online]Available:https://www.nextmsc.com/report/lpwa-market

[3] K. Mekki et al., “A comparative study of LPWAN technologies for large-scale IoT deployment,” ICT Express, vol. 5, no. 1, pp. 1–7, Mar. 2019, doi: 10.1016/j.icte.2017.12.005.

[4] Statista, “IoT connected devices worldwide 2019–2030,” [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[5] J. Guo et al., "IEEE 802.19.3 Coexistence Recommendations for IEEE 802.11 and IEEE 802.15.4 Based Systems Operating in SUB-1 GHz Frequency Bands," in IEEE Communications Standards Magazine, vol. 7, no. 2, pp. 72-82, June 2023,doi: 10.1109/MCOMSTD.0009.2100046.

[6] Keysight Technologies, “QualNet Network Simulator,” [Online]. Available: https://www.keysight.com/jp/ja/assets/3122-1395/technical-overviews/QualNet-Network-Simulator.pdf

[7] J. Guo et al., “Impact of network profiles on 802.11ah and IEEE 802.15.4g coexistence performance,” IEEE 802.19 Working Group contribution, doc. IEEE 802.19-19/0070r1, 2019.

[8] "IEEE Standard for Information technology--Telecommunications and information exchange between systems - Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Sub 1 GHz License Exempt Operation," in *IEEE Std 802.11ah-2016 (Amendment to IEEE Std 802.11-2016, as amended by IEEE Std 802.11ai-2016)* , vol., no., pp.1-594, 5 May 2017, doi: 10.1109/IEEESTD.2017.7920364.

[9] S.-J. Kim, M. Aono, and M. Hara, “Tug-of-war model for the two-bandit problem:Nonlocally-correlated parallel exploration via resource conservation,” BioSystems, vol. 101, no. 1, pp. 29–36, Jul. 2010, doi: 10.1016/j.biosystems.2010.04.002.

[10] S.-J. Kim, M. Aono, and E. Nameda, “Efficient decision-making by volume-conserving physical object,” New J. Phys., vol. 17, no. 8, Art. no. 083023, Aug. 2015,doi: 10.1088/1367-2630/17/8/083023.

[11] S.-J. Kim et al., “Decision maker based on atomic switches,” AIMS Mater. Sci., vol. 3, no. 1, pp. 245–259, 2016, doi: 10.3934/matersci.2016.1.245.

[12] S.-J. Kim and M. Aono, “Amoeba-inspired algorithm for cognitive medium access,” Nonlinear Theory and Its Applications, IEICE, vol. 5, no. 2, pp. 198–209, Apr. 2014, doi: 10.1587/nolta.5.198.

[13] J. Ma et al., “A Reinforcement-Learning-Based Distributed Resource Selection Algorithm for Massive IoT,” Applied Sciences, vol.9. no. 18, 3730, Sep. 2019, doi: 10.3390/app9183730.

[14] D. Yamamoto et al., “Performance evaluation of reinforcement learning based distributed channel selection algorithm in massive IoT networks,” IEEE Access, vol. 10, pp. 67870–67882, 2022, doi: 10.1109/ACCESS.2022.3186703.

[15] I. Urabe et al., “Combinatorial MAB-Based Joint Channel and Spreading Factor Selection for LoRa Devices,” *Sensors*, vol. 23, no. 15, p. 6687, 2023, doi: 10.3390/s23156687.

[16] R. Nomura et al., “Experimental Evaluation of Reinforcement Learning-Based Power Consumption Reduction Method for Bluetooth Low Energy,” *IEICE Technical Report*, vol. 124, no. 362, pp. 47–52, 2025.