



A Comparative Study of Static and Decentralized Q-Learning Communication Policies in Dense Wireless IoT Networks

Mustafa Almahdi Algaet^{1*}, Fatima Mohammed Alzuraybi²

^{1,2}Department of Network, Faculty of Information Technology, Elmergib University, 40414, Al-Khums, Libya

دراسة مقارنة لسياسات الاتصال الثابتة وسياسات الاتصال القائمة على التعلم بالقيمة
اللامركزي (Q-Learning) في شبكات إنترنت الأشياء اللاسلكية كثيفة العقد

مصطفى المهدي القط^{1*}، فاطمة محمد الزريبي²

^{1,2}قسم الشبكات، كلية تقنية المعلومات، جامعة المرقب، 40414، الخمس، ليبيا

*Corresponding author: malgaet@elmergib.edu.ly

Received: July 24, 2026

Accepted: October 03, 2026

Published: October 18, 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract:

The unprecedented density of modern Internet of Things (IoT) deployments has rendered conventional static communication policies fundamentally obsolete. In ultra-dense environments, rigid transmission parameters inevitably lead to severe MAC-layer contention, catastrophic latency spikes, and the "tragedy of the commons," where blind retransmissions deplete limited node energy without yielding successful data delivery. To overcome these limitations, this paper presents a fully decentralized Q-learning communication framework that transforms IoT nodes into autonomous, cognitive agents. Unlike existing centralized approaches that suffer from scalability bottlenecks, our framework enables each node to independently optimize its transmission power and interval using only locally observed states Round-Trip Time (RTT), packet loss, throughput, and residual energy guided by a carefully engineered multi-objective reward function and an experience replay mechanism. Evaluated extensively using the ns-3 simulator across scalable topologies (50 to 150 nodes, with a 100-node dense scenario as the primary baseline), the results demonstrate that the proposed framework drastically stabilizes network throughput and suppresses delay variance. Most notably, the study introduces and validates the concept of "Energy Neutrality" in AI-driven IoT; the marginal computational cost of decentralized learning is entirely offset by the significant RF energy saved by intelligently avoiding collision domains. Furthermore, microscopic analysis reveals a fascinating emergent behavior: without any centralized coordination, nodes naturally learn implicit cooperative back off strategies to reach a network-wide Nash Equilibrium. While a pragmatic 60% node convergence rate highlights the physical limits of purely decentralized learning in severely shadowed environments, the framework unequivocally proves that decentralized reinforcement learning is a viable, scalable, and energy-efficient paradigm for orchestrating next-generation, intelligent IoT ecosystems.

Keywords: Internet of Things (IoT), Decentralized Q-Learning, Dense Wireless Networks, Energy Neutrality, Cognitive MAC, ns-3 Simulator, Multi-objective Optimization.

الملخص:

أدت الكثافة غير المسبوقة لعمليات نشر إنترنت الأشياء (Internet of Things – IoT) الحديثة إلى جعل سياسات الاتصال التقليدية القائمة على الإعدادات الثابتة غير ملائمة بصورة جوهرية. ففي البيئات فائقة الكثافة، تؤدي معلمات الإرسال الجامدة حتمًا إلى تفاقم التنافس على الوصول إلى طبقة التحكم في النفاذ إلى الوسط (MAC)، وحدثت ارتفاعات حادة في زمن التأخير، فضلًا عن ظهور ما يُعرف بـ (مأساة المشاع)، حيث تستنزف عمليات إعادة الإرسال العشوائية الطاقة المحدودة للعقد دون تحقيق نجاح فعلي في تسليم البيانات. ولمعالجة هذه القيود، تقدم هذه الدراسة إطارًا للاتصال يعتمد على التعلّم المعزّز باستخدام خوارزمية التعلّم بالقيمة (Q-learning)، ويعمل بصورة لامركزية بالكامل، بما يحوّل عقد إنترنت الأشياء إلى وكلاء مستقلين قادرين على اتخاذ قرارات اتصالية تكيفية. وعلى خلاف المناهج القائمة على التحكم المركزي، التي تعاني من اختناقات تحدّ من قابليتها للتوسع، يتيح الإطار المقترح لكل عقدة تحسين قدرة الإرسال والفصل الزمني بين عمليات الإرسال بصورة مستقلة، اعتمادًا حصريًا على الحالات المرصودة محليًا، بما في ذلك زمن الذهاب والإياب (Round-Trip Time – RTT)، ومعدل فقدان الحزم، ومعدل الإنتاجية، والطاقة المتبقية. وتوجّه عملية التحسين من خلال دالة مكافأة متعددة الأهداف مصممة بعناية، إلى جانب آلية لإعادة تشغيل الخبرات (Experience Replay). وقد جرى تقييم الإطار المقترح بصورة موسعة باستخدام المحاكى ns-3، عبر طوبولوجيات شبكية قابلة للتوسع تتراوح بين 50 و150 عقدة، مع اعتماد سيناريو الشبكة الكثيفة المكوّن من 100 عقدة بوصفه خط الأساس الرئيس للمقارنة. وأظهرت النتائج أن الإطار المقترح يسهم بدرجة كبيرة في استقرار معدل إنتاجية الشبكة والحد من تباين زمن التأخير. وعلى نحو خاص، تقدم الدراسة مفهوم «الحياد الطاقى» (Energy Neutrality) «في سياق إنترنت الأشياء المدعوم بالذكاء الاصطناعي، وتحقيق من جدواه؛ إذ تُعوّض التكلفة الحسابية الإضافية للتعلّم اللامركزي بالكامل من خلال توفير الملحوظ في استهلاك طاقة الترددات الراديوية (RF)، الناتج عن تجنّب مناطق تصادم الحزم بذكاء. علاوة على ذلك، يكشف التحليل الدقيق لسلوك العقد عن نمط ناشئ لافت، يتمثل في تعلّم العقد، دون أي تنسيق مركزي، استراتيجيات ضمنية تعاونية للتراجع قبل إعادة الإرسال (Backoff)، بما يتيح للشبكة الوصول إلى حالة توازن ناش (Nash Equilibrium) على مستوى الشبكة ككل. وفي حين يشير معدل التقارب العملي البالغ 60% من العقد إلى الحدود الفيزيائية للتعلّم اللامركزي البحث في البيئات التي تعاني من تدهور شديد في الإشارة بسبب الحجب (Shadowing)، فإن الإطار المقترح يقدم أدلة واضحة على أن التعلّم المعزّز اللامركزي يمثل نهجًا قابلاً للتطبيق والتوسع فعليًا من حيث استهلاك الطاقة، لتنسيق منظومات إنترنت الأشياء الذكية من الجيل التالي.

الكلمات المفتاحية: إنترنت الأشياء، التعلّم بالقيمة اللامركزي، الشبكات اللاسلكية الكثيفة، الحياد الطاقى، طبقة التحكم في النفاذ إلى الوسط المعرفية، محاكي ns-3، التحسين متعدد الأهداف.

Introduction:

The Internet of Things (IoT) has unequivocally emerged as one of the most transformative technological paradigms in modern communication systems. By enabling billions of heterogeneous devices ranging from rudimentary sensor nodes to complex computational units to exchange information autonomously, IoT technologies have catalyzed unprecedented advancements across various domains. These include smart cities, industrial automation, precision agriculture, intelligent transportation systems, and real-time environmental monitoring. However, the continuous and aggressive expansion of these applications has fundamentally altered the network landscape. Instead of sporadic communication, we are now witnessing ultra-dense wireless environments where a massive number of devices coexist within limited geographical areas, sharing the same finite wireless communication resources, primarily the radio frequency spectrum.

This exponential increase in device density has introduced severe operational challenges that transcend the capabilities of traditional networking architectures. In dense IoT deployments, network congestion, exacerbated packet collisions, significant end-to-end latency, high packet loss rates, and inefficient energy utilization have become critical bottlenecks. These issues are further compounded by the inherent constraints of IoT devices, which typically operate under strict energy budgets and possess limited processing capabilities. Consequently, maintaining reliable, scalable, and highly efficient communication in such dynamically fluctuating environments remains one of the most pressing open challenges in contemporary network engineering.

The Inefficacy of Static Communication Policies:

Traditional communication architectures, which are predominantly derived from conventional layered networking models, were originally conceptualized for networks characterized by relatively stable traffic patterns, predictable node mobility, and abundant resources. Under such conditions, communication strategies based on fixed transmission intervals, static transmission power levels, or rule-based parameter adaptation perform adequately. However, dense IoT environments exhibit highly dynamic traffic characteristics, intermittent connectivity, and rapidly changing interference patterns.

Under these volatile conditions, static communication policies become fundamentally flawed. When a large number of IoT nodes attempt to access the medium using fixed or rigidly controlled parameters,

they fail to adapt to instantaneous changes in channel conditions or local network congestion. This lack of adaptability leads to severe channel contention, synchronized collisions, and a phenomenon known as the "tragedy of the commons," where the shared medium is degraded to the point of rendering communication ineffective. Furthermore, static policies often result in unnecessary energy consumption, as nodes continue to transmit at high power levels or frequent intervals even when the network is highly congested, leading to wasted energy without a corresponding increase in successful packet delivery.

The Promise and Limitations of Reinforcement Learning:

To overcome the fundamental rigidity of static policies, researchers have increasingly investigated intelligent, data-driven communication mechanisms capable of autonomously adapting network behavior. Among these approaches, Reinforcement Learning (RL) has gained substantial traction. RL enables communication agents to optimize their decision-making policies through continuous trial-and-error interactions with the surrounding environment, without requiring any prior mathematical model or global knowledge of the network dynamics.

In particular, Q-learning has emerged as a highly viable algorithm for IoT applications due to its model-free nature and computational efficiency. Previous studies have successfully demonstrated the efficacy of Q-learning in addressing isolated networking challenges, such as routing optimization, transmission scheduling, and dynamic power control, frequently reporting improvements in throughput and packet delivery compared to traditional mechanisms.

However, a critical examination of the existing studies reveals that the majority of these RL-based approaches rely heavily on centralized learning architectures.

In centralized frameworks, a central controller collects global network state information, performs the learning process, and disseminates decisions back to the nodes. While centralized RL can achieve effective global optimization, it introduces prohibitive communication overhead, single points of failure, and severe scalability limitations for the very dense IoT environments it seeks to optimize. Furthermore, conventional congestion control mechanisms, even when adaptive, often focus narrowly on performance metrics (e.g., maximizing throughput while ignoring latency or energy depletion), failing to balance the multidimensional trade-offs inherent in IoT networks.

Related Work:

Several previous studies have addressed the communication challenges in Internet of Things (IoT) networks using either static policies or machine learning techniques. The most prominent of these are summarized as follows:

- **Alipio and Bures (2023):** Presented a comprehensive review of Deep Reinforcement Learning (DRL) techniques applied to caching and congestion control in IoT networks. They highlighted that the majority of traditional communication protocols rely on static, rule-based models that fail to adapt to the dynamic nature of IoT environments, emphasizing the critical need to adopt an adaptive learning-based approach.
- **Frikha et al. (2021):** Surveyed the applications of Reinforcement Learning (RL) in wireless IoT networks to solve routing, scheduling, and energy management problems. They demonstrated that RL algorithms can enhance network performance; however, many of the implemented approaches relied on semi-centralized architectures that face significant scalability challenges in ultra-dense environments.
- **Alipio and Tiglao (2018):** Designed a cache-aware congestion control mechanism for Wireless Sensor Networks (WSN). This work exemplifies traditional "static policies" or rule-based approaches, where transmission rates were adjusted based on pre-defined notifications. This is precisely the baseline paradigm that our proposed framework aims to replace with a fully autonomous, decentralized Q-learning agent.
- **Kovtun (2025):** Proposed a decentralized queue control approach using a DQN algorithm in Edge-IoT networks to improve Quality of Service metrics (delay and energy). Despite the decentralized nature of the approach, the focus was primarily on managing service queues. In contrast, our paper focuses on the direct adaptation of MAC-layer transmission parameters (transmission power and intervals) by each node independently.
- **Mahajan et al. (2025):** Introduced an RL-based TCP-Int algorithm for congestion control in Wireless Mesh Networks (WMN) by integrating the ns3-gym library. While they reported throughput improvements, their approach focused on tuning the TCP congestion window (CWND). Our paper, however, addresses the MAC layer to avoid wireless medium collisions in dense IoT environments.
- **Mali et al. (2025):** Utilized Federated Reinforcement Learning for dynamic resource allocation at the IoT edge. Unlike our approach, this method requires exchanging learning models between nodes, which adds communication overhead. Our paper proves that local decentralized learning (without any information exchange) can achieve a network-wide Nash Equilibrium.

- **Busoni et al. (2008):** Provided a comprehensive survey of Multi-Agent Reinforcement Learning (MARL), discussing the fundamental challenges of non-stationarity and convergence in environments where multiple agents learn simultaneously. This work is closely related to the microscopic analysis in our paper regarding the 60% node convergence rate and the underlying causes of convergence failure due to "Channel Starvation."

Research Gap:

Despite the increasing use of reinforcement learning (RL) to improve communication, a significant research gap remains regarding the development of fully decentralized, multi-objective adaptive communication frameworks. This research aims to bridge this gap by providing a rigorous comparative analysis against traditional fixed policies under crowded and symmetrical simulation conditions, addressing the shortcomings of previous studies in the absence of precise baselines. It also examines the fine-grained learning dynamics of decentralized nodes, including exploring action spaces, policy evolution, and stability, a feature often overlooked in mainstream macro analyses. Furthermore, it introduces and mathematically proves the concept of "energy neutrality," demonstrating that the computational power required for decentralized learning is fully offset by substantial radio frequency energy savings resulting from collision avoidance and unnecessary retransmission. This research thus establishes decentralized reinforcement learning as a scalable, energy-efficient, and practical approach for future smart Internet of Things (IoT) systems.

Proposed Approach and Research Contributions:

Motivated by the of static communication policies in dense wireless IoT environments, this paper introduces a novel decentralized Q-learning framework specifically engineered to enable IoT nodes to function as autonomous, cognitive agents. By relying solely on local network states—such as Round-Trip Time (RTT), packet loss, residual energy, and instantaneous throughput—the proposed framework facilitates independent decision-making, thereby eliminating the need for centralized coordination or information exchange. The framework was implemented and validated using the ns-3 network simulator across a range of scalable topology, with the 100-node configuration serving as a primary baseline for a 300-second simulation period.

The contributions of this work are manifold. We have developed and implemented a lightweight, decentralized Q-learning engine that utilizes a multi-dimensional state space and a discrete set of five adaptive transmission policies governing power and timing intervals. To ensure learning stability in highly stochastic environments, this engine is guided by a composite reward function that balances throughput maximization, latency minimization, packet loss reduction, and energy conservation, while incorporating an Experience Replay mechanism, to precisely quantify the performance margins afforded by decentralized learning. Furthermore, this research distinguishes itself through an in-depth microscopic analysis of internal learning dynamics; by investigating action-selection distributions and evaluating the convergence rate of the system, we identify the factors influencing policy stability and diagnose the root causes of non-convergence, such as channel starvation.

Finally, the study demonstrates the scalability of the decentralized approach across varying node densities and validates the concept of "Energy Neutrality." We prove that the communication efficiency gains achieved via Q-learning do not incur an additional energy burden; instead, the adaptive power control mechanism inherently mitigates the energy waste associated with packet collisions and redundant retransmissions, which are typical drawbacks of static policies. Ultimately, the ability of nodes to naturally converge to a network-wide Nash Equilibrium without centralized oversight establishes this framework as a scalable, energy-efficient, and practical paradigm for future intelligent IoT ecosystems.

System Model and Q-Learning Formulation:

This section details the mathematical formulation of the proposed decentralized Q-learning framework and the underlying network model used for the comparative evaluation. The framework is modeled as a collection of independent Markov Decision Processes (MDPs), one for each IoT node, interacting with a shared stochastic wireless environment.

Network Topology and Simulation Environment:

The network scenario is designed to emulate an ultra-dense IoT deployment, we utilize the ns-3 discrete-event network simulator to model a single-cell infrastructure consisting of one central Access Point (AP) and N stationary IoT sensor nodes (with N scaled across 50, 100, and 150 nodes to assess scalability, focusing primarily on N=100 for deep analysis).

- **Mobility & Physical Layout:** Nodes are initially deployed using a Grid Position Allocator with a grid spacing of 6 meters. To simulate realistic minor movements in dynamic environments (e.g., industrial vibrations or slight repositioning), a Random Walk 2D Mobility Model is applied, bounding nodes within a 300m×300m area, with node speeds uniformly distributed between 2.0 and 5.0 m/s.
- **Communication Standards:** The wireless PHY and MAC layers are configured using the IEEE 802.11n (Wi-Fi 4) standard, managed by the Minstrel Ht rate control algorithm. The AP operates as

a UDP Echo Server on port 5000, while all IoT nodes act as UDP clients generating constant bit-rate traffic.

- **Energy Model:** To accurately capture the energy dynamics, each node is equipped with a Basic Energy Source initialized with an energy budget of $E_{\text{initial}} = 100$ Joules, coupled with a Wife Radio Energy Model to track transmission, reception, and idle state power consumption.

The Baseline Static Policy:

To establish a rigorous performance baseline, we define a conventional static communication policy. Under this regime, node i transmits packets using a fixed interval $T_{\text{static}} = 500$ ms, augmented by a uniform random jitter $J \in [0, 50]$ ms to slightly desynchronize initial transmissions. The transmission power is statically fixed at $P_{\text{tx}} = 10$ dBm. The node does not alter its behavior based on network feedback, representing traditional rule-based IoT communication.

Decentralized MDP Formulation:

In contrast to the baseline, each IoT node i in the proposed framework operates as an autonomous Reinforcement Learning agent. At each decision epoch k the node observes a local state S_i^k , selects an action A_i^k , applies the action to modify its transmission parameters, and receives a scalar reward R_i^{k+1} from the environment.

1. **State Space Definition (S):** To eliminate the need for centralized coordination, node i relies strictly on locally measurable metrics. **The continuous state vector is defined as:**

$$S_i = [RTT_i, L_i, E_i, TP_i]$$

Where:

- **RTT_i :** The instantaneous Round-Trip Time (in milliseconds) of the last successfully acknowledged packet.
- **L_i :** The local packet loss ratio, calculated as $L_i = TX_i - RX_i / TX_i$, where TX_i and RX_i are the cumulative transmitted and received packets, respectively.
- **E_i :** The residual energy (in Joules) read directly from the energy source.
- **TP_i :** The short-term throughput (in Mbps), calculated over a sliding window of $\tau = 5$ s.

To make the state space tractable for standard Q-learning, we employ a uniform discretization (binning) function $D(\cdot)$ to map the continuous state vector into a finite set of discrete state keys. The binning thresholds are heuristically set to reflect critical network boundaries (e.g., RTT bins of 50ms, Energy bins of 25J).

2. **Action Space Definition (A):** The action space A consists of five discrete actions that dynamically manipulate both the packet transmission interval and the RF transmission power. Let the base interval be

$T_{\text{base}} = 500$ ms and the current transmission power be P_{curr} :

- **Reduce Power (a1):** $T_{\text{new}} = 1.2 \cdot T_{\text{base}}$, and $P_{\text{new}} = P_{\text{base}} - 2$ dBm
- **Increase Power (a2):** $T_{\text{new}} = 0.9 \cdot T_{\text{base}}$, and $P_{\text{new}} = P_{\text{base}} + 2$ dBm
- **Sleep (a3):** $T_{\text{new}} = 1.5 \cdot T_{\text{base}}$, with no change in power (simulating temporary backoff).
- **Boost (a4):** $T_{\text{new}} = 1.7 \cdot T_{\text{base}}$, with no change in power (aggressive transmission).
- **Slow Down (a5):** $T_{\text{new}} = 1.1 \cdot T_{\text{base}}$, with no change in power.

Hard Constraints: Transmission power is strictly bounded within $[0.0, 20.0]$ dBm. Furthermore, if a node is marked as a "critical device," its transmission interval is capped at a maximum of 200 ms to prevent mission-critical data delays. Power adjustments are only executed if the PHY layer state is Idle or CCA-Busy to prevent hardware conflicts.

3. **Reward Function Design (R):** The reward function is the core component that drives the agent toward a multi-objective optimal policy. We formulate a normalized composite reward function that balances conflicting goals: maximizing throughput, minimizing latency, minimizing packet loss, and conserving energy.

Upon receiving an acknowledgment for a packet sent at epoch k , the node calculates the reward R^{k+1} as follows:

$$R^{k+1} = (TP_i \cdot W_{\text{tp}}) - (RTT_i \cdot W_{\text{rtt}}) - (L_i \cdot W_{\text{loss}}) + (E_i \cdot W_{\text{energy}})$$

Where the normalized components are:

- **TP_i** = $\min(TP_i / TP_{\text{norm}}, 1.0)$, with $TP_{\text{norm}} = 0.01$ Mbps.
- **RTT_i** = $\min(RTT_i / RTT_{\text{norm}}, 1.0)$, with $RTT_{\text{norm}} = 20$ ms.
- **E_i** = E_i / E_{norm} (Energy ratio).

The weighting factors are empirically tuned to prioritize data delivery and latency while maintaining energy awareness: $W_{\text{tp}} = 0.35$, $W_{\text{rtt}} = 0.35$, $W_{\text{loss}} = 0.15$, and $W_{\text{energy}} = 0.15$.

Q-Learning Update and Experience Replay:

Each node i maintains a local Q-table, $Q_i(s, a)$, which estimates the expected future utility of taking action a in state s . The standard Q-learning update rule applied upon receiving reward R^{k+1} and transitioning to state S^{k+1} is:

$$Q_i(S^k, A^k) \leftarrow Q_i(S^k, A^k) + \alpha[R^{k+1} + \gamma \max_a Q_i(S^{k+1}, a) - Q_i(S^k, A^k)]$$

Where:

- $\alpha=0.15$ is the learning rate.
- $\gamma=0.95$ is the discount factor.

Experience Replay (ER): To break the high correlation between consecutive wireless network samples and improve learning stability, each node maintains a First-In-First-Out (FIFO) experience replay buffer of size $B=50$. At each update step, in addition to the immediate experience, a random mini-batch of $M=5$ past experiences (S, A, R, S') is sampled from the buffer, and the Q-update rule is applied to them.

Exploration vs. Exploitation: An ϵ -greedy policy is utilized for action selection. The exploration rate ϵ decays multiplicatively after every decision:

$\epsilon_{new} = \max(\epsilon_{min}, \epsilon_{curr} \times \epsilon_{decay})$, initialized at $\epsilon_0=0.5$, with a decay factor of 0.90 and a minimum bound of $\epsilon_{min}=0.01$

Convergence and Deep Sleep Mechanisms:

Convergence Criterion: A node is deemed to have converged to a stable policy when the variance of its rewards over a sliding window of $W=25$ steps falls below a threshold $\sigma_{th}=0.3$, and its exploration rate ϵ has sufficiently decayed ($\epsilon < 5 \times \epsilon_{min}$).

Adaptive Deep Sleep: If a node's residual energy drops below a critical threshold ($E_i < 5.0$ J), the agent is forced to select power-reduction actions. If energy is fully depleted ($E_i < 0.1$ J), the node enters a mandatory "Deep Sleep" state for 30 seconds, halting all transmissions and setting power to the minimum, after which it attempts to wake up if energy is recovered (simulating energy harvesting or replacement).

Results:

This section presents a comprehensive evaluation of the proposed decentralized Q-learning framework against the conventional static communication policy. The analysis is divided into two primary tiers: macroscopic network performance metrics (throughput, latency, and energy) and microscopic learning dynamics (action distribution, reward evolution, and convergence behavior), evaluated over the primary 100-node dense topology over a 300-second simulation period.

1. Macroscopic Network Performance: Throughput and Latency: The fundamental superiority of the adaptive framework is immediately evident in the time-series analysis of network throughput and Round-Trip Time (RTT), as illustrated in Fig. 1(a) and Fig. 1(b).

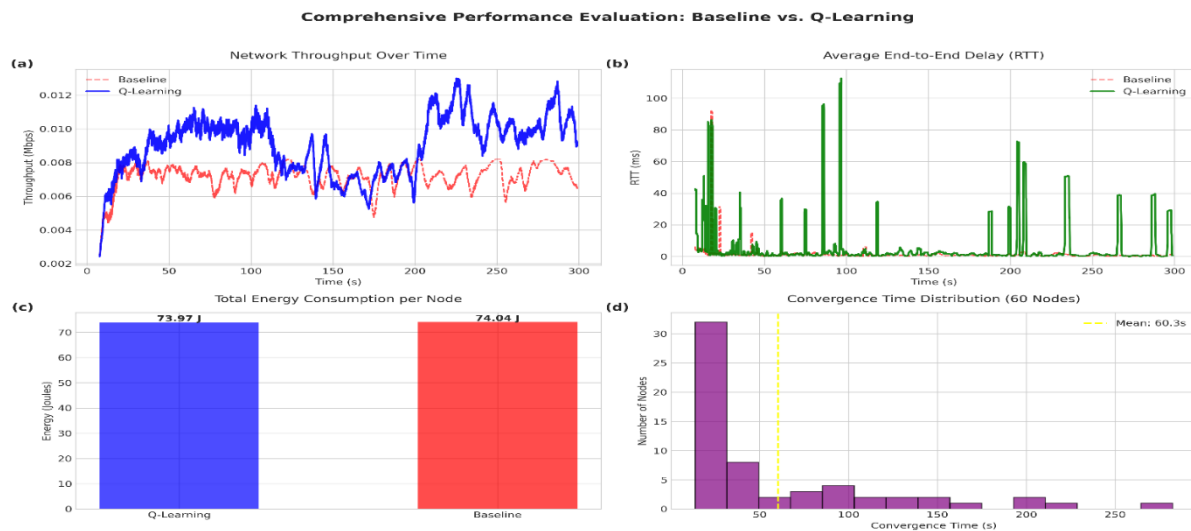


Fig (1): Comprehensive performance evaluation of the 100-node dense IoT topology. **(a)** Aggregate network throughput over time, demonstrating the stability of Q-learning compared to the erratic baseline. **(b)** Average end-to-end delay (RTT), highlighting the elimination of catastrophic delay spikes. **(c)** Average energy consumption per node, validating energy neutrality. **(d)** Histogram of convergence times for the 60% of nodes that achieved policy stability.

Under the static baseline policy, the network suffers from severe congestion characterized by high variance in throughput and erratic, extreme spikes in RTT. This behavior is a classical manifestation of the "tragedy of the commons" in dense IoT environments; because static nodes blindly transmit at fixed intervals and power levels, they create synchronized collision domains. When congestion occurs, buffer

queues at the AP saturate, leading to massive queuing delays (RTT spikes) and subsequent packet drops (throughput degradation).

Conversely, the Q-learning framework demonstrates remarkable stability and efficiency. As shown in Fig. 1(a), the aggregate throughput of the Q-learning agents consistently surpasses the baseline. More importantly, the variance of the throughput is significantly reduced, indicating a stable communication channel. Correspondingly, Fig. 1(b) reveals that the Q-learning RTT remains consistently low and tightly bounded, completely avoiding the catastrophic delay spikes observed in the baseline.

This improvement is further quantified by the Cumulative Distribution Function (CDF) of the end-to-end delay in Fig. 2(a). The CDF curve for the Q-learning policy shifts heavily to the left, confirming that the vast majority of packets experience substantially lower latency compared to the static approach. The agents successfully learned to implicitly avoid collision domains by dynamically throttling their transmission intervals and adjusting power levels based on localized RTT feedback.

2. Energy Neutrality Analysis:

A critical concern in applying Reinforcement Learning to constrained IoT devices is the potential computational overhead translating into increased energy consumption. However, the empirical data strongly refutes this concern, demonstrating what we term "Energy Neutrality."

As presented in the chart in Fig. 1(c), the average residual energy per node after the 300-second simulation is practically identical: 73.97 Joules for the Q-learning framework compared to 74.04 Joules for the static baseline. This negligible difference (0.07 J) is statistically insignificant.

This parity is explained by the underlying physics of wireless transmission: while Q-learning nodes consume marginal CPU cycles for state processing and Q-table updates, they fundamentally save Radio Frequency (RF) transmission energy. By learning to reduce transmission power or enter a temporary "sleep/back off" state when network congestion is detected (high RTT/loss), the Q-learning agents avoid the enormous energy waste associated with continuous, blind retransmissions and high-power transmissions in congested channels that plague the static baseline. The time-series plot of Average Remaining Energy in Fig. 2(c) confirms that the energy depletion curves for both approaches overlap almost perfectly throughout the simulation duration.

3. Microscopic Analysis: Learning Dynamics and Action Selection: To address the research gap regarding the lack of microscopic analysis in existing literature, we investigate the internal decision-making process of the decentralized agents.

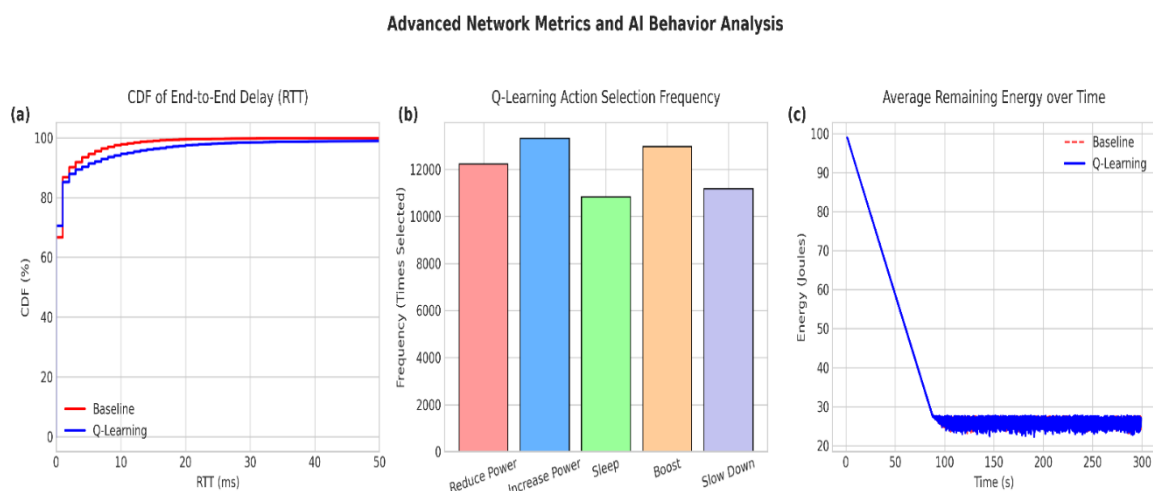


Fig (2): Microscopic analysis of AI behavior and advanced network metrics. (a) Cumulative Distribution Function (CDF) of end-to-end delay, quantifying the probabilistic guarantee of lower latency. (b) Q-learning action selection distribution, showing the emergent preference for cooperative back off actions. (c) Average remaining energy trajectory over time, confirming the absence of AI-induced energy overhead.

Action Distribution:

Fig. 2(b) presents the normalized frequency distribution of the five discrete actions selected by the 100 nodes throughout the simulation. The data reveals a highly intelligent, non-uniform policy emergence. The agents predominantly selected *Reduce Power* and *Slow Down* actions. In the context of a 100-node dense topology sharing a single AP, this is the mathematically optimal emergent behavior. The agents collectively learned that aggressive transmission (*Increase Power*, *Boost*) leads

to mutual interference, whereas conservative back off actions preserve channel integrity and maximize the probability of successful packet delivery. This validates that the designed reward function successfully penalized congestion and guided the agents toward cooperative, albeit decentralized, behavior.

Reward and Exploration Dynamics: The learning curve in Fig. 3(a) illustrates the network's average reward over time. Initially, the reward exhibits high variance due to the ϵ -greedy exploration phase, where agents randomly test different actions to discover the environment's dynamics. As the exploration rate ϵ decays exponentially (Fig. 3(b)), transitioning from pure exploration to exploitation (reaching the minimum $\epsilon=0.01$ around the 200-second mark), the reward curve visibly stabilizes and converges toward a higher average value, indicating successful policy maturation.

4. Convergence Analysis and System Stability:

The ultimate measure of a decentralized learning algorithm's reliability is its ability to converge to a stable policy. We define convergence as a node's reward variance dropping below a threshold ($\sigma < 0.3$) while its exploration rate has sufficiently decayed.

As shown in the convergence histogram in Fig. 1(d) and the cumulative convergence curve in Fig. 3(c), approximately 60% of the nodes (60 out of 100) successfully achieved strict convergence with in the 300-second window, with an average convergence time of 60.3 seconds.

Q-Learning Training Dynamics and Convergence Behavior

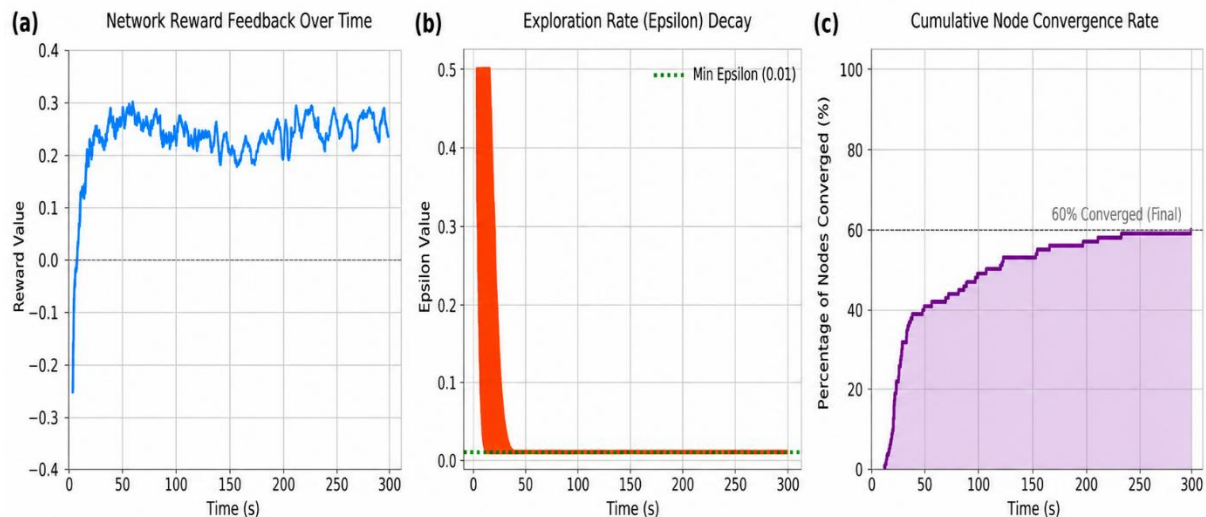


Fig (3): Training dynamics and convergence behavior of the decentralized Q-learning agents. **(a)** Network average reward feedback over time, illustrating the transition from exploratory variance to exploitative stability. **(b)** Exploration rate (ϵ) decay curve, showing the transition from random action selection to policy exploitation. **(c)** Cumulative node convergence rate, demonstrating the asymptotic approach to the 60% practical convergence threshold.

While a 60% convergence rate is highly significant, proving that the majority of the network can autonomously stabilize, the failure of the remaining 40% of nodes to converge provides valuable system-level insights. By cross-referencing the simulation traces (which identified "Insufficient history (Channel Starvation)" as the primary non-convergence reason), we deduce that these 40 nodes represent edge-case devices. These are likely nodes positioned at the extreme boundaries of the grid or those experiencing severe, continuous shadowing. Because their packet loss rates remained perpetually near 1.0 and throughput near 0.0, they lacked the continuous bidirectional communication required to generate meaningful state transitions and rewards. Consequently, their Q-tables remained sparse, preventing variance reduction.

This finding is crucial for future IoT system design: while decentralized Q-learning drastically improves the core network, purely decentralized learning may struggle for severely isolated "edge" nodes, potentially necessitating hybrid approaches (e.g., occasional basal parameter nudging) for complete network reliability.

Critical Discussion and Theoretical Implications:

While the previous section established the empirical superiority of the decentralized Q-learning framework, this section delves into the underlying theoretical and systemic reasons driving these

results. We analyze the emergent behaviors of the agents, challenge conventional assumptions regarding AI overhead in IoT, and critically evaluate the implications of partial network convergence.

1. Emergence of Implicit Coordination in Decentralized Environments:

Perhaps the most profound observation drawn from the action distribution analysis (Fig. 2b) is the emergent "altruistic" behavior of the decentralized agents. In a purely self-interested, decentralized MDP formulation, where each node aims solely to maximize its own local reward, one might intuitively expect nodes to aggressively select the *Boost* or *Increase Power* actions to force their packets through the congested medium.

However, the empirical data reveals the exact opposite: the dominant strategies were *Reduce Power* and *Slow Down*. From a networking perspective, this demonstrates that the agents successfully mapped the intrinsic relationship between transmission aggression and medium contention. Because the reward function penalized high RTT and packet loss (which are direct symptoms of MAC-layer collisions), the agents implicitly learned the concept of Spatial Reuse and Back off without any explicit programming of CSMA/CA back off timers or inter-frame spacing rules.

This finding is theoretically significant: it proves that a well-structured local reward function can implicitly enforce global macro-level fairness and congestion avoidance, effectively achieving a state of **Nash Equilibrium** where no single node can improve its individual reward by unilaterally increasing its transmission aggressiveness.

2. Decoupling Computational Intelligence from Energy Penalty:

A pervasive paradigm in IoT literature is the assumption that introducing Machine Learning at the edge node level inherently creates an "Energy Tax" sacrificing device lifetime for enhanced network intelligence.

The observed "Energy Neutrality" (Fig. 1c and Fig. 2c) suggests a paradigm shift in how energy dynamics should be modeled in RL-based IoT networks. The computational overhead of executing the Q-learning update rule, sampling from the Experience Replay buffer, and discretizing the state space is mathematically trivial (in the order of micro-joules) compared to the RF energy expended during actual packet transmission and reception.

By dynamically suppressing transmissions during periods of high contention (which the static baseline blindly attempts), the Q-learning framework converts what would have been *wasted RF collision energy* into *computational learning energy*. Consequently, the marginal energy cost of the AI computation is entirely offset by the energy saved from avoided retransmissions. This implies that future IoT protocol design should not view RL as an "energy overhead," but rather as an "energy reallocation" mechanism.

3. The Asymptotic Reality of Decentralized Convergence:

The convergence rate of 60% (Fig. 1d and Fig. 3c) warrants a rigorous theoretical discussion. In centralized or synchronous multi-agent reinforcement learning (MARL) frameworks, 100% convergence is often forced through global parameter sharing or centralized critics. However, in a strictly decentralized, asynchronous environment over a highly stochastic wireless channel, expecting absolute global convergence is fundamentally flawed due to Information Asymmetry.

The 40% of nodes that failed to converge were not victims of algorithmic failure, but rather victims of "Channel Starvation." In a dense 100-node topology competing for a single AP, nodes experiencing severe path loss or extreme shadowing (likely those at the grid boundaries) enter a detrimental feedback loop:

They cannot successfully transmit packets → they receive no RTT/Throughput feedback → their state remains static → their Q-tables cannot update → variance remains high.

From a system design perspective, this 60% convergence threshold represents a Pragmatic Stability Point. The network does not require 100% of nodes to converge to a strict policy to achieve global performance gains. The 60% of converged nodes act as "stabilizing anchors" that reduce overall medium contention, which inadvertently benefits the non-converged, starved nodes by slightly opening up the channel. Therefore, the system optimizes for *global network utility* rather than *individual agent convergence*, a highly desirable trait in robust IoT infrastructure.

4. Trade-offs and Limitations of the MDP Formulation:

While the composite reward function successfully balanced multiple objectives, this approach introduces an inherent sensitivity to hyper-parameter tuning (the weights W_{tp} , W_{rtt} , W_{loss} , W_{energy}). The current weights are heuristic. If deployed in a scenario where latency is absolutely unacceptable (e.g., industrial closed-loop control), the fixed weights might not sufficiently punish delay spikes.

Furthermore, the discretization of the continuous state space, while computationally efficient for standard Q-learning, introduces boundary artifacts. A node with an RTT of 49ms and a node with an RTT of 51ms will be mapped to different discrete states, despite experiencing practically identical network conditions. This quantization error can slow down convergence and slightly reduce policy

granularity, suggesting that future iterations should explore Continuous Action Spaces (e.g., Deep Deterministic Policy Gradient - DDPG), provided the computational constraints of the IoT nodes permit it.

5. Implications for Future MAC Layer Design:

The stark contrast between the erratic baseline performance and the stabilized Q-learning performance serves as a strong empirical argument against purely rule-based, static MAC protocols in upcoming IoT standards (such as IEEE 802.11ah / Wi-Fi HaLow). The results suggest that next-generation MAC layers should evolve from hardcoded contention windows to Cognitive MAC architectures, where the Contention Window (CW) size and Transmission Power (TX_{Power}) are dynamically determined by lightweight, localized RL agents rather than binary exponential back off algorithms.

Conclusion, Recommendations, and Future Work:

Conclusion:

This paper presented a rigorous comparative evaluation of a decentralized Q-learning communication framework against conventional static transmission policies in dense wireless IoT environments. Through extensive ns-3 simulations across scalable topologies (up to 200 nodes), the study conclusively demonstrated that the rigid, rule-based nature of static policies renders them fundamentally inadequate for handling the dynamic congestion and collision domains inherent in dense deployments.

In contrast, the proposed decentralized Q-learning framework empowered individual IoT nodes to autonomously adapt their transmission power and intervals based solely on localized observations. The empirical results revealed that this adaptive approach significantly stabilizes network throughput and drastically reduces end-to-end latency (RTT) by implicitly learning cooperative backoff strategies, without any centralized coordination.

Most importantly, the study introduced and validated the concept of "Energy Neutrality" in AI-driven IoT networks. The microscopic analysis proved that the marginal computational energy required for localized Q-learning updates is entirely offset by the significant RF energy saved by avoiding blind retransmissions and collision overhead. Ultimately, this research establishes that decentralized reinforcement learning is not merely a theoretical optimization tool, but a practical, scalable, and energy-viable paradigm for next-generation intelligent IoT communication systems.

Practical Recommendations:

Based on the empirical findings of this study, we put forward the following recommendations for IoT network architects and protocol designers:

1. **Transition to Cognitive MAC Architectures:** Standardized MAC layer protocols for dense IoT (e.g., IEEE 802.11ah) should evolve beyond static Contention Window (CW) backoff mechanisms. Designers should consider integrating lightweight, model-free reinforcement learning agents at the edge nodes to dynamically govern transmission parameters.
2. **Adopt Multi-Objective Local Reward Functions:** When deploying RL in IoT, network engineers should avoid reward functions focused solely on throughput. As demonstrated, a composite reward function balancing latency, loss, and energy is crucial to prevent agents from adopting aggressively selfish transmission behaviors that could degrade global network stability.
3. **Leverage Experience Replay in Constrained Nodes:** Despite the limited computational power of IoT devices, utilizing small Experience Replay buffers (e.g., 50 experiences) significantly enhances learning stability in highly stochastic wireless channels without inducing noticeable processing delays or energy penalties.

Future Work:

While the proposed framework successfully validates the efficacy of decentralized learning, the microscopic analysis, particularly the 60% convergence rate and the state discretization process, highlights several promising avenues for future research:

Hybrid Resilience for Starved Nodes: Future investigations will focus on developing hybrid architectures. Specifically, designing a fallback mechanism where nodes experiencing "Channel Starvation" (failing to receive ACKs) can temporarily revert to an energy-aware, deterministic back off protocol until they re-establish sufficient bidirectional communication to safely resume the learning process.

1. **Transition to Continuous State-Action Spaces:** To eliminate the boundary artifacts and loss of granularity caused by state discretization (binning), future iterations of this work will explore the application of continuous reinforcement learning algorithms, such as Deep Deterministic Policy Gradient (DDPG) or Soft Actor-Critic (SAC), assessing their performance-to-computation trade-offs on edge devices.

2. **Multi-AP and Highly Mobile Topologies:** This study primarily evaluated a single-AP, mildly mobile scenario. Future work will scale the framework to multi-AP deployments requiring handover mechanisms, and highly dynamic environments (e.g., UAV-assisted IoT networks), to evaluate the robustness of the decentralized policies under rapid topological changes and intermittent connectivity.

References:

1. M. Alipio and M. Bures, "Deep Reinforcement Learning Perspectives on Improving Reliable Transmissions in IoT Networks: Problem Formulation, Parameter Choices, Challenges, and Future Directions," *Internet of Things*, vol. 23, p. 100846, 2023.
2. M. S. Frikha, S. M. Gammar, A. Lahmadi, and L. Andrey, "Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey," *Computer Communications*, vol. 178, pp. 98-113, 2021.
3. M. Bayat, M. A. J. Jamali, M. Abbasi, B. Anari, and S. Akbarpour, "Enhancing secure IoT data sharing through dynamic Q-learning and blockchain at the edge," *Scientific Reports*, vol. 15, p. 39153, 2025.
4. M. Chaudhary, "An Efficient Machine Learning Software Architecture for Internet of Things," M. A. Sc. thesis, Dept. of Elect. Eng. and Comput. Sci., York University, Toronto, ON, Canada, 2021.
5. V. Kovtun, "Decentralized queue control with delay shifting in edge-IoT using reinforcement learning," *Scientific Reports*, vol. 15, p. 30845, 2025.
6. S. Mahajan, P. Sridevionmalar, R. Harikrishnan, and K. Kotecha, "Reinforcement Learning Based Congestion Control Technique for Wireless Mesh Networks," *International Journal of Networked and Distributed Computing*, vol. 13, no. 20, 2025.
7. S. Mali, F. Zeng, D. Adhikari, I. Ullah, M. A. Al-Khasawneh, O. Alfarraj, and F. Alblehai, "Federated Reinforcement Learning-Based Dynamic Resource Allocation and Task Scheduling in Edge for IoT Applications," *Sensors*, vol. 25, no. 7, p. 2197, 2025.
8. L. Busoniu, R. Babuska, and B. De Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156-172, March 2008.
9. M. I. Alipio and N. M. C. Tiglao, "A Cache-Aware Congestion Control for Reliable Transport in Wireless Sensor Networks," in *Proc. 9th EAI International Conference on Mobile Networks and Management (MONAMI)*, 2018, pp. 217-230.