The OSCAR PUBLISHING Services
American Journal of Applied Science and Technology

VOLUME Vol.06 Issue 04 2026
PAGE NO. 185-190
DOI 10.37547/ajast/Volume06Issue04-25

Analysis and Halon-IOT Concept to Reduce Data Transmission Delays in IOT Networks

Kholiqulov Sherzod
Associate Professor of the Department of "Communication and Information Technologies" of the University of Tashkent, Uzbekistan

Makhmudov Doniyorbek Ravshanovich
Master's Student of UMFT University, Uzbekistan

Received: 18 March 2026; **Accepted:** 06 April 2026; **Published:** 30 April 2026

Abstract: This article analyzes published scientific works on reducing data transmission delays in IoT networks. Within the framework of the analysis, resource allocation approaches based on edge-cloud continuum, fog-edge architecture, partial offloading, serverless offloading, SLA-aware scheduling, and deep reinforcement learning were compared. Based on the selected works, the article proposes a new HALON-IoT (Hybrid Adaptive Latency-Optimized Network for IoT) conceptual model. This model combines mist/edge/fog/cloud layers to provide ultra-low latency for critical flows and elastic scalability for analytical tasks. The main innovation of the proposed model is the integration of packet-level prioritization, partial offloading, SLA-risk scoring, and SDN/DRL orchestration into a single control contour. The article is written in a conceptual format and requires experimental validation based on emulation or a real test environment at the next stage.

Keywords: IoT, latency, hybrid network, edge-cloud, fog computing, partial offloading, DRL, SLA-aware scheduling, SDN.

INTRODUCTION:

IoT ecosystems are rapidly expanding in industries such as industry, healthcare, smart cities, transportation, and agriculture. However, due to the increasing data flows, device heterogeneity, and services requiring near-real-time responses, traditional cloud-only architectures are no longer sufficient in many cases. In particular, transmission delay, queuing delay, and processing delay in the chain from sensors to gateways, from gateways to edge/fogs, and from there to the cloud increase the overall end-to-end latency.

In recent years, researchers have been turning to hybrid network models to address this problem. Such models distribute computing resources across mist, edge, fog, and cloud layers rather than a single central cloud. As a result, time-sensitive packets are processed closer to the source, while workloads that require large volumes or historical analysis are offloaded to the cloud. However, simply adding layers to the architecture is not enough: it is also important to consider which layer to route tasks to, at what time, and over which network path.

The purpose of this paper is twofold: first, to systematically review Q1 papers that focus on reducing IoT latency; second, to propose a new conceptual hybrid model based on this analysis. The paper is not a fully experimental work; it combines elements of a review and conceptual design. Therefore, the new model in it should be tested in a testbed, simulation, or pilot environment before being implemented in practice.

RESEARCH METHODOLOGY AND SELECTED SOURCES

The literature selection was based on the following criteria:

(a) the article must have been published between 2024-2026;

(b) the journal falls into category Q1;

(c) the central problem of the work is related to latency, scheduling, offloading, or resource allocation in an IoT or edge-cloud continuum environment;

(d) the article contains innovation at the architectural or algorithmic level.

The articles were selected from the following publications: Journal of Network and Computer Applications, Future Generation Computer Systems, Computer Communications, Digital Communications and Networks, Journal of Cloud Computing, Scientific Reports, and IEEE Internet of Things Journal.

The analysis was conducted based on the same criteria: architecture type, offloading method, latency reduction mechanism, decision-making algorithm, SLA/QoS considerations, energy and resource efficiency, and practical constraints. This approach allowed for a comparison of different cases within the same framework.

Table 1. Comparative analysis of models in IoT networks

Ref.	Source	Model type	Main idea	Conclusion on the delay
[1]	JNCA, 2025	Partial offloading + matching theory	It divides tasks into subtasks and assigns them to fog/VMs based on deadline and processing efficiency.	Reasonable service delay; removes tasks that cannot be completed within the deadline early.
[2]	FGCS, 2024	QoS-aware serverless offloading	A two-stage utility-driven serverless policy in the cloud-to-edge continuum.	It improves edge proximity deployment by considering QoS classes and budget.
[3]	ComCom, 2024	Latency-cost-sensitive placement	Optimizes the placement of service modules in a hierarchical fog.	Reduces latency caused by communication hops and placement errors.
[5]	J Cloud Comp, 2025	DQN-PPO-GNN-RL	Manages scheduling and resource allocation in dynamic workloads with AI.	Minimum latency 14-20 ms; job execution accuracy 98-99%.
[6]	Sci Rep, 2025	FPA-TS metaheuristic	It optimizes makespan, cost, and utilization together for the edge-cloud continuum.	Beats benchmarks in cost and makespan; edge cost 41.76 (1000 tasks).
[7]	Sci Rep, 2025	Hybrid fog-edge IoMT	Real-time processing between edge and fog in health monitoring.	500-750 ms vs 1400-1600 ms; threat detection 30 ms.
[8]	Sci Rep, 2026	AICDQN	Priority-aware task offloading based on GRU-LSTM + D4QN.	Delay -33.39%; energy efficiency +57.74%; task drop -81.25%.
[9]	Sci Rep, 2026	SLA-DRL	Scheduler based on SLA tier, risk score and action pruning.	SLA violation -41.6%; average latency -32.1%; 63.9 ms latency.

RESULTS AND DISCUSSION

The first is to move away from full offloading and toward partial offloading. A 2025 article in the Journal of Network and Computer Applications proposes dividing a task into pieces and assigning them to fog nodes and VMs based on deadlines and processing efficiency. The authors demonstrate early filtering of useless tasks, i.e. tasks that will not be completed within the deadline, using the RABP and TAC algorithms. This approach reduces latency because a large task is not completely moved to the cloud, but useful subtasks are processed in layers closer to the source.

Second is QoS-aware and utility-driven offloading. The 2024 work in Future Generation Computer Systems proposes a two-stage serverless offloading policy for the cloud-to-edge continuum. The solution aims to maximize utility for different service classes while operating under resource budget constraints. This result shows that latency reduction can be

achieved not only by the minimum time principle, but also by separating quality of service classes. Especially in event-driven IoT services, it is easier to control P95/P99 latency by placing serverless functions close to the edge.

Third, consider placement and routing costs together. A 2024 paper in Computer Communications optimizes application placement in hierarchical fog computing environments in terms of latency and communication cost. This is important because in many systems, even if the scheduler works well, misplaced service modules add unnecessary hops across the network. As a result, even though the computation is fast, the end-to-end response time remains high.

Fourth, the integration of edge node selection and task scheduling. A 2025 Digital Communications and Networks paper explores the combination of edge node selection and task scheduling in latency-sensitive monitoring and control for industrial IoT. This approach aims to simultaneously achieve fault tolerance and low latency in manufacturing processes. This idea is particularly relevant in practice for industrial control loops and video inspection workflows.

The fifth trend is resource pre-allocation using AI/DRL. A 2025 paper in the Journal of Cloud Computing proposes a cloud-edge hybrid model based on a combination of DQN-PPO-GNN-RL. The authors report 95-98% task scheduling accuracy, 94-97% resource allocation accuracy, 98-99% job execution accuracy, and minimal latency in the range of 14-20 ms in high-load, dynamic workload, and latency-sensitive application scenarios. This result implies that latency can be significantly reduced not only by increasing resources, but also by graph-aware and policy-driven management.

The sixth trend is the parallel development of metaheuristic and DRL approaches. The 2025 FPA-TS algorithm in Scientific Reports simultaneously optimizes makespan, execution cost, and resource utilization for the edge-cloud continuum. It reduces the edge execution cost to 41.76 in 1000 task instances, and performs 10-26% better than competing algorithms. This means that while reducing latency, it is possible to balance system cost and resource utilization.

The seventh trend is domain-specific hybrid architectures. The 2025 IoMT fog-edge model in Scientific Reports showed real-time health monitoring processing times in the range of 500-750 ms, significantly faster than the 1400-1600 ms in the cloud-only model. Also, the threat detection time was reduced to 30 ms. These figures show that latency is directly related to network topology and task layering.

The eighth trend is the enhancement of next-generation DRL schedulers with SLA and priority awareness. In a 2026 Scientific Reports paper, the AICDQN model shows a 33.39% delay reduction, a 57.74% energy efficiency improvement, and an 81.25% task drop rate reduction. In another 2026 paper, an SLA-aware DRL approach reduced SLA violations by 41.6%, average latency by 32.1%, and energy consumption by 28.5%; the average latency reached 63.9 ms and the task completion rate reached 94.1%. These results indicate that future IoT schedulers should integrate deadline, priority, and violation risk.

However, the analysis also shows that most work focuses on compute scheduling, but does not sufficiently expose deep integration with network path selection, congestion-aware routing, packet prioritization, and SDN management. Many models perform well in simulation, but do not fully demonstrate how they perform in real-world network situations such as packet loss, jitter, intermittent connectivity, gateway failure, or LPWAN/5G/Wi-Fi heterogeneity.

Proposed model

Based on the above analysis, this paper proposes the concept of HALON-IoT (Hybrid Adaptive Latency-Optimized Network for IoT). The main idea of the model is to reduce latency not by a single layer or a single algorithm, but by integrated management at three levels:

(1) initial filtering of the data stream and classification of packets into classes;

(2) flexible choice of computing placement;

(3) Dynamic management of network routing and forwarding policies.

The architecture consists of four layers. The fog/device layer performs event-based filtering, compression, and priority marking of the stream coming from the sensor. The edge layer (gateway or local micro-edge) performs critical tasks that need to respond within 20-50 ms: anomaly detection, actuator trigger, initial image inference, local cache, and session state. The regional fog layer performs more cooperative computing: multi-camera fusion, predictive analytics, near-batch stream processing, queue balancing. The cloud layer handles model training, historical analytics, archiving, and global policy updates.

In HALON-IoT, all tasks are divided into three classes: C1 - ultra-critical (deadline < 20 ms), C2 - interactive (20-100 ms), C3 - analytical/batch (>100 ms). C1 tasks are processed by default at the edge layer; if the edge queue exceeds the threshold, only compressed or feature-level partial offloading is performed to the regional fog . In C2 tasks, an adaptive choice is made between edge and fog. C3 tasks are transmitted to the cloud, but if the backhaul is busy, temporary buffering and pre-aggregation are used in the fog.

A single cost function is proposed for decision making:

J = w1*D + w2*Q + w3*B + w4*E + w5*S.

Here D is the predicted end-to-end delay, Q is the queue length or queue saturation, B is the link occupancy and packet loss risk, E is the energy consumption, and S is the SLA violation risk. For each task, the scheduler dynamically selects the appropriate w coefficients depending on the deadline and priority. Thus, the model does not just select the 'nearest node', but also the safest and most time-efficient transmission-computing combination.

In the control plane, the SDN controller and the DRL orchestrator work together. The SDN layer manages paths, queue policies, and packet classes; the DRL orchestrator selects task placement, autoscaling, and partial offloading. This combined approach fills a major gap in the existing literature, where routing and scheduling are typically optimized separately. HALON-IoT integrates them into a single decision loop.

The expected benefits are: first, tail latency is reduced for critical flows; second, the amount of unnecessary raw data transmitted to the cloud is reduced; third, edge bottleneck is mitigated using regional fog; fourth, deadline miss rate is reduced due to SLA-based prioritization; fifth, the trade-off between energy and network consumption is manageable. However, there are also issues such as controller overhead, state synchronization, and telemetry accuracy when implementing the model.

From a practical perspective, the HALON-IoT model is suitable for applications such as agricultural monitoring, video analytics, industrial lines, health monitoring, and smart transportation. For example, in agro-IoT systems, soil sensor, drone video, and weather station data can be distributed in different layers: signal or threshold violation is performed at the edge, multi-sensor analysis is performed in the fog, and long-term yield modeling is performed in the cloud. This approach reduces the load on the backhaul and speeds up operational decisions.

However, several issues remain open for future research. The first is to jointly optimize routing, scheduling, and security policies. The second is to verify the model's adaptability in different transport environments such as 5G, Wi-Fi 6, LPWAN, and satellite backhaul. The third is to combine it with federated or split learning approaches to maintain high accuracy without sending the data entirely to the cloud. The fourth is to evaluate the impact of jitter, node failure, and mobility on tail latency based on a real testbed.

Therefore, the literature suggests that the most promising direction is a combination of multi-layered hybrid architecture and intelligent orchestration. Rather than relying solely on the edge or solely on the cloud, hybrid continuum and policy-aware management are more likely to deliver sustained latency reduction.

Criteria for the experiment

For the next stage of experimental evaluation of the HALON-IoT model, a specific set of KPIs should be defined. First of all, indicators such as median latency, P95 and P99 tail latency, deadline miss ratio, task completion rate, queue waiting time, packet loss, jitter and backhaul utilization should be measured. Drawing conclusions based on average latency alone is not enough, because in real IoT systems, the user or management contour is often violated precisely

due to tail latency. Therefore, when evaluating the model, it is recommended to maintain separate KPI profiles for critical tasks, separate for interactive tasks and separate for analytics tasks.

As an experimental platform, environments such as iFogSim2, EdgeCloudSim, YAFS or RegionalEdgeSimPy can be used. It is advisable to divide the test scenarios into at least four types:

(1) normal loading;

(2) burst traffic;

(3) edge node failure or gateway overload;

(4) link degradation or intermittent connectivity.

In each scenario, the HALON-IoT model should be compared with cloud-only, edge-only, simple round-robin scheduling, QoS-only heuristic, and non-SLA DRL as a baseline. This way, it becomes clear under which conditions the proposed model has the advantage.

Telemetry collection is also important in practical implementation. If the scheduler operates with the wrong state, even the best DRL model will make the wrong decisions. Therefore, parameters such as queue metrics, CPU/RAM usage, wireless signal quality, per-flow RTT, packet retransmission, energy profile, and SLA violation history must be collected in near real-time intervals. In addition, the control-plane overhead itself should not increase latency; that is, monitoring and control traffic should not crowd out useful traffic. Only if these aspects are taken into account can the HALON-IoT concept be transformed from a scientific hypothesis into a practical, stable architecture.

CONCLUSION

The article analyzed Q1 papers on reducing latency in IoT networks and drew three main conclusions. First, partial offloading, deadline-aware scheduling, and edge-fog-cloud continuum architectures are significantly superior to the cloud-only approach. Second, DRL, metaheuristic, and SLA-aware methods better manage the tradeoff between latency, energy, and resource consumption. Third, the most powerful future solutions will be hybrid models that manage routing, placement, and scheduling together. On this basis, the HALON-IoT concept was proposed. The scientific value of this model lies in the fact that it combines the network path, task priority, partial offloading, and computation layers into a single decision mechanism. In the future, it is advisable to emulate this model in environments such as iFogSim/EdgeCloudSim/RegionalEdgeSimPy and test it in real pilot systems.

REFERENCES

1. Sushma, SA, Madhunisha, E., Addya, SK, Rahman, S., Pal, S., Karmakar, C. Delay-aware partial task offloading using multicriteria decision model in IoT-fog-cloud networks. Journal of Network and Computer Applications, 242, 104278, 2025. https://doi.org/10.1016/j.jnca.2025.104278
2. Russo Russo, G., Ferrarelli, D., Pasquali, D., Cardellini, V., Lo Presti, F. QoS-aware offloading policies for serverless functions in the Cloud-to-Edge continuum. Future Generation Computer Systems, 156, 1-15, 2024. https://doi.org/10.1016/j.future.2024.02.019
3. Oliveira, LT, Bittencourt, LF, Genez, TAL, de Lara, E., Peixoto, MLM Enhancing modular application placement in a hierarchical fog computing: A latency and communication cost-sensitive approach. Computer Communications, 216, 95-111, 2024. https://doi.org/10.1016/j.comcom.2024.01.002
4. Kumari, M., Singh, MP, Singh, AK A latency sensitive and agile IIoT architecture with optimized edge node selection and task scheduling. Digital Communications and Networks, 2025. https://doi.org/10.1016/j.dcan.2025.04.012
5. Lilhore, UK, Simaiya, S., Sharma, YK et al. Cloud-edge hybrid deep learning framework for scalable IoT resource optimization. Journal of Cloud Computing, 14, 5, 2025. https://doi.org/10.1186/s13677-025-00729-w
6. Dankolo, NM, Radzi, NHM, Mustaffa, NH et al. Optimizing resource allocation for IoT applications in the edge cloud continuum using hybrid metaheuristic algorithms. Scientific Reports, 15, 14409, 2025. https://doi.org/10.1038/s41598-025-97648-2
7. Islam, U., Alatawi, MN, Alqazzaz, A. et al. A hybrid fog-edge computing architecture for real-time

health monitoring in IoMT systems with optimized latency and threat resilience. Scientific Reports, 15, 25655, 2025. https://doi.org/10.1038/s41598-025-09696-3

8. Anand, J., Karthikeyan, B. Adaptive and intelligent customized deep Q-network for energy-efficient task offloading in mobile edge computing environments. Scientific Reports, 16, 5456, 2026. https://doi.org/10.1038/s41598-025-34765-y
9. Yamsani, N., P, CR SLA aware deep reinforcement learning for adaptive EdgeCloud task scheduling. Scientific Reports, 16, 10037, 2026. https://doi.org/10.1038/s41598-026-40237-8
10. Fan, Q., Bai, J., Zhang, H., Yi, Y., Liu, L. Delay-aware resource allocation in fog-assisted IoT networks through reinforcement learning. IEEE Internet of Things Journal, 9(7), 5189-5199, 2022. https://doi.org/10.1109/JIOT.2021.3111079