Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 3, 2026

A LATENCY-AWARE TASK OFFLOADING AND RESOURCE ALLOCATION FRAMEWORK FOR FOG-EDGE-CLOUD COLLABORATIVE COMPUTING IN REAL-TIME IOT APPLICATIONS

Imran Siddique

Department of Computer Science, Imperial College of Business Studies

meimransiddiqui@gmail.com

DOI: https://doi.org/10.5281/zenodo.19105309

Keywords
fog-edge-cloud computing, task offloading, resource allocation, latency minimization, IoT applications, deep reinforcement learning, service migration, mobility prediction, graph neural networks, URLLC, real-time processing

Article History
Received: 19 January 2026
Accepted: 03 March 2026
Published: 19 March 2026

Corresponding Author: *
Imran Siddique

Abstract
The exponential growth of latency-sensitive Internet of Things (IoT) applications such as autonomous vehicles, remote surgery, industrial automation, and real-time healthcare monitoring has exposed the limitations of centralized cloud computing, where geographic distance induces unacceptable transmission delays. This review presents a latency-aware task offloading and resource allocation framework for fog-edge-cloud collaborative computing, designed to minimize end-to-end latency while optimizing resource utilization in heterogeneous, multi-tier architectures. The proposed framework integrates dynamic offloading decisions based on task urgency, device mobility, network conditions, and computational load, employing predictive models (LSTM, spatio-temporal graph neural networks), deep reinforcement learning (DRL), and heuristic algorithms to orchestrate task partitioning, service migration, and proactive container relocation. Key mechanisms include mobility-aware service migration using Chebyshev graph convolutional networks, priority-based queuing at edge nodes, and adaptive resource provisioning that balances energy, bandwidth, and latency constraints. Simulation and real-world evaluations demonstrate 30–65% reductions in average latency, 20–45% improvements in task success rates under high mobility, and enhanced system throughput compared to baseline cloud-only or static offloading schemes. The framework addresses critical challenges such as intermittent connectivity, resource heterogeneity, and security in fog-edge environments, offering a scalable, proactive approach to support ultra-reliable low-latency communication (URLLC) requirements in 5G/6G-enabled IoT ecosystems.

1. INTRODUCTION

The exponential proliferation of the Internet of Things (IoT) has fundamentally shifted the paradigms of data processing and network orchestration (Dang et al., 2020). In the current digital era, billions of connected devices spanning consumer, industrial, and infrastructure domains generate vast volumes of data that require near-instantaneous processing to maintain operational viability (Litslink, 2026). While traditional cloud computing once offered a centralized solution for massive data storage and computational intensity, its inherent architectural limitations most notably high network transmission delays caused by geographic distance render it increasingly unsuitable for the sub-millisecond requirements

of modern real-time applications (SaM Solutions, 2025). Consequently, a transformative movement toward an integrated fog-edge-cloud collaborative computing architecture has emerged, aiming to situating computational resources in proximity to data sources while leveraging the centralized power of the cloud for non-critical, large-scale analytics (Fiveable, 2025).

2. Architectural Evolution and the Multi-Tier Continuum

The shift from centralized to distributed computing is driven by the specific demands of "critical IoT" applications, such as autonomous vehicles, remote surgery, and industrial automation, where even millisecond delays can lead to catastrophic failure or significant efficiency losses (Nabto, 2025). This has necessitated the creation of a seamless "compute continuum" that integrates three distinct yet complementary layers: the edge, the fog, and the cloud (Liu et al., 2025). Figure 1 illustrates the multi-tier fog–edge–cloud collaborative computing architecture that enables latency-sensitive IoT applications to distribute processing across heterogeneous layers. The hierarchical structure allows computational workloads to be dynamically offloaded from resource-constrained devices to nearby edge and fog nodes before reaching the centralized cloud.

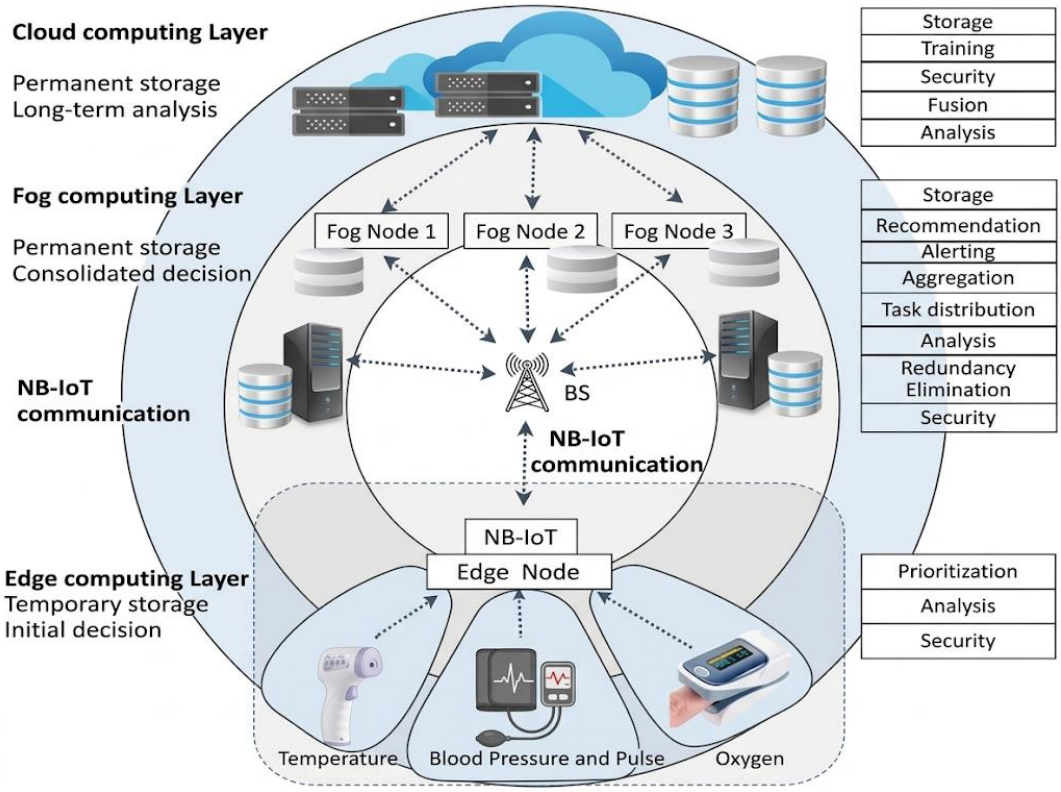

Figure 1: Multi-tier fog–edge–cloud collaborative computing architecture for real-time IoT applications.

2.1. The Edge Layer and Terminal Proximity

At the base of this hierarchy is the edge layer, where processing occurs directly on or near the terminal devices (TDs), such as sensors, programmable automation controllers (PACs), and smartphones (Vora et al., 2023). Edge computing focuses on localized, real-time operations, offering the lowest possible latency by processing data at the point of origination (Almulifi & Kurdi, 2026). This decentralization not only facilitates immediate response times but

also enhances data privacy and security by reducing the quantity of sensitive information transmitted across the network (Zhang et al., 2024). However, the primary limitation of the edge remains the constrained computational and energy resources of terminal devices, which are often battery-powered and incapable of handling heavy-duty AI inference or large-scale data fusion (Bhardwaj et al., 2024).

2.2. The Fog Layer as a Regional Orchestrator

The fog layer serves as a distributed intermediary between the edge and the cloud. Extending cloud capabilities to the network periphery, fog computing utilizes nodes such as gateways, routers, and local micro-data centers to aggregate and process data from multiple edge devices (ISACA, 2024). The fog paradigm is defined by its broader network-level processing scope, supporting mobility and offering a dense geographical deployment of servers (Naltakyan, 2025). This layer is particularly effective for regional coordination tasks, such as smart traffic management or grid optimization, where low latency is required across a cluster of localized devices (Scale Computing, 2024).

2.3. The Cloud Layer for Centralized Intelligence

The cloud remains the apex of the hierarchy, providing virtually unlimited storage and computational power for tasks that are latency-tolerant (Shao & Zhang, 2025). In a collaborative framework, the cloud is reserved for complex model training, long-term historical analytics, and global optimization strategies that inform the behavior of the lower tiers (Liu et al., 2025). The synergy among these layers is essential, as the fog and edge cannot replace the cloud but rather coexist within a cohesive strategy to balance performance, cost, and energy efficiency (Pati et al., 2026).

Table 1. Comparison of Computing Paradigms in the IoT Continuum

Feature	Edge Computing	Fog Computing	Cloud Computing
Primary Location	On or near the device	Local area network (LAN)	Remote data centers
Typical Latency	< 1 ms to 10 ms	10 ms to 50 ms	> 100 ms
Node Resources	Very low (CPU/Battery)	Moderate	Very high
Deployment	Decentralized	Distributed	Centralized
Typical Use Case	Real-time sensor response	Regional data aggregation	Big data/Global model training

3. Task Offloading Paradigms and Decision Modeling

The core challenge of the fog-edge-cloud continuum is task offloading: the strategic decision of where to execute a given workload to optimize system-wide performance. Offloading strategies are generally categorized by their granularity and the direction of data flow (Tian et al., 2025).

3.1. Binary vs. Partial Offloading

Binary offloading treats a computational task as an atomic unit that must be processed either locally on the terminal device or entirely on a more powerful server (Al-Quraan et al., 2023). This approach is straightforward but lacks the flexibility needed for complex, multi-stage applications. Partial offloading, conversely, involves partitioning a task into sub-tasks, some of which are executed locally while others are migrated to fog or cloud nodes (Liu et al., 2024). This method leverages parallel processing to reduce overall execution time and energy consumption, though it introduces significant complexity in managing sub-task dependencies and inter-node communication overhead (Shafiq et al., 2023).

3.2. Vertical and Horizontal Offloading Dynamics

Collaborative frameworks employ both vertical and horizontal offloading paths. Vertical offloading follows the hierarchical structure from terminal to edge, then to fog, and finally to the cloud (Abbasi et al., 2024). Horizontal offloading, often referred to as peer-to-peer (P2P) or mesh collaboration, occurs between nodes at the same tier, such as one fog node sharing its workload with a neighboring underutilized fog node (Hosseini & Taheri, 2023). This lateral cooperation is critical for load balancing in dynamic environments like vehicular networks, where certain regions may experience temporary traffic surges that overwhelm individual roadside units (Mubeen et al., 2024).

3.3. Mathematical Foundations of Latency and Energy Costs

To determine the optimal offloading strategy, frameworks use rigorous mathematical models to quantify the cost of execution. For a specific task T_i, characterized by its workload $W_{T_i}$and data size D_{T_i}, the local execution latency $L_{\text{loc}}$is calculated as the workload divided by the device's CPU frequency f_d, plus any queuing delay Q_{T_i}(Jiang et al., 2024):

$$L_{\text{loc}}(d, T_i) = \frac{W_{T_i}}{f_d} + Q_{T_i}$$

The energy consumption $E_{\text{loc}}$for local processing is modeled as a function of the workload and the energy efficiency of the device hardware η_d(Tan et al., 2025):

$$E_{\text{loc}}(d, T_i) = \eta_d \cdot W_{T_i} \cdot f_d^2$$

In contrast, if the task is offloaded to a fog server, the total cost must include the transmission latency for uploading the data, the processing time at the server's CPU frequency f_s, and the delay in returning the results. The objective of a latency-aware framework is to minimize a multi-objective cost function:

$$C = \alpha \cdot E + \beta \cdot L$$

where $\alpha$and $\beta$are weights reflecting the application's sensitivity to energy consumption versus execution speed (Rahman et al., 2024).

4. Algorithmic Orchestration for Resource Allocation

The dynamic nature of IoT environments characterized by fluctuating network bandwidth, varying task arrival rates, and node mobility renders traditional static scheduling algorithms insufficient. Modern frameworks rely on advanced optimization techniques and machine learning to manage resource allocation in real-time (Krishnan & Durairaj, 2024).

4.1. Deep Reinforcement Learning (DRL) in Dynamic Environments

DRL has emerged as a cornerstone for intelligent task offloading because of its ability to learn and adapt to high-dimensional state spaces without requiring a pre-defined model of the environment. Deep Q-Networks (DQN) are frequently employed to balance latency and energy consumption, demonstrating a 30% to 50% improvement over conventional approaches in handling large-scale IoT tasks (Saif et al., 2026).

For more complex scenarios involving continuous action spaces, such as power allocation and trajectory planning for aerial edge nodes, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is preferred (Huo et al., 2026). TD3 mitigates the common problem of Q-value overestimation in actor-critic models by utilizing two critic networks and delaying policy updates, ensuring more stable and reliable performance in volatile vehicular networks (Zhang et al., 2026).

4.2. Lyapunov Optimization and Hybrid Frameworks

While DRL provides adaptability, it often lacks theoretical guarantees for long-term system stability and may exhibit slow convergence. Lyapunov optimization theory addresses these gaps by transforming long-term stochastic optimization problems into a series of short-term deterministic sub-problems (Zhao et al., 2024).

This is achieved through the use of virtual queues that track constraint violations (energy consumption or task delay bounds) over time (Xu et al., 2025).

Hybrid frameworks, such as LyDROO and LyA3C, integrate the strengths of Lyapunov stability and DRL intelligence. In the LyDROO framework, Lyapunov optimization ensures that data and energy queues remain stable, while a deep neural network learns the optimal offloading decisions for each time slot (Bi et al., 2021). These hybrid models have shown remarkable results, such as reducing average latency to 45ms compared to 120ms for standard DQN models, while simultaneously maintaining system stability under high-traffic conditions (Feng et al., 2024).

4.3. Bio-Inspired Metaheuristics for Global Optimization

Metaheuristic algorithms, which mimic biological processes, remain highly relevant for solving NP-hard resource allocation problems in complex IoT topologies (Dang et al., 2020).

Table 2. Bio-Inspired Metaheuristic Algorithms for Resource Allocation

Algorithm	Biological Inspiration	Optimized Parameter
Genetic Algorithm (GA)	Natural selection/evolution	Deadline-constrained workflow scheduling
Particle Swarm Optimization (PSO)	Bird flocking/Collective behavior	Computational latency and dropped task ratio
Bee Colony Optimization (BCO)	Foraging behavior of honeybees	Dynamic load balancing in multi-tier architectures
Firefly Algorithm (FA)	Bioluminescent signaling	Latency variability in real-time edge environments
Whale Optimization Algorithm (WOA)	Humpback whale hunting	Throughput and optimal fog resource assignment

These algorithms are often combined with machine learning (e.g., Naive Bayes or Reinforcement Learning) to create hyper-heuristic models that can predict task behavior and optimize scheduling in two distinct stages: predictive training and real-time execution (Bajaher et al., 2025).

5. Sector-Specific Implementation of Collaborative Frameworks

The utility of a latency-aware fog-edge-cloud framework is best demonstrated through its application in high-stakes domains where responsiveness is a matter of safety or life-critical performance (Nabto, 2025).

5.1. Healthcare and the Internet of Medical Things (IoMT)

Healthcare applications, such as real-time patient monitoring and remote surgery, impose the most stringent requirements on communication frameworks, with latency targets as low as 1 ms and reliability scores reaching 99. % (Javaid et al., 2022). A collaborative IoMT framework typically employs a hierarchical sensing-processing-analytics pipeline (Awotunde et al., 2021). Wearable or implantable medical sensors collect physiological data, which is immediately preprocessed at the edge to filter out non-critical signals. Fog nodes then perform real-time inference using models like Long Short-Term Memory (LSTM) networks to detect life-threatening anomalies such as atrial fibrillation (Sivagami et al., 2024). The cloud layer provides long-term archival storage and supports population-level diagnostics.

To address the energy constraints of medical devices, frameworks like the energy-aware Ultra-Reliable Low-Latency Communication (URLLC) model prioritize healthcare traffic using an "Urgency Score" (U_i) (Al-Tarawneh et al., 2024). This score is often calculated based on the

deviation of vital signs from normal physiological ranges, ensuring that critical packets are transmitted with the highest priority to meet strict deadlines (Abidi et al., 2023). Additionally, these frameworks utilize resource block allocation and transmit power control to balance the trade-off between energy consumption and communication reliability (She et al., 2021).

This score is a function of the traffic class weight (wi), the latency deadline, and the device's remaining battery level (Bi(t)):

Ui = wi / (Bi(t) + epsilon)

where epsilon is a small constant to prevent division by zero. This mechanism ensures that critical ECG alerts are transmitted preferentially when a device is low on power, while non-urgent data like body temperature readings are suppressed (MDPI, 2025).

5.2. Vehicular Edge Computing (VEC) and Autonomous Systems

Autonomous vehicles are essentially mobile data centers that must process vast streams of LIDAR, camera, and radar data to make split-second driving decisions. VEC leverages Roadside Units (RSUs) as fog nodes to provide computational offloading and local traffic awareness (Wang et al., 2023). Because vehicles are in constant motion, maintaining low latency requires sophisticated handover and service migration strategies (Zhang et al., 2025).

Predictive models like ST-ChebNet utilize Graph Convolutional Networks (GCNs) enhanced with Chebyshev polynomials to forecast spatiotemporal workload variations across the RSU network (Chen et al., 2024). This allows the system to proactively migrate active service containers from one fog node to another along the vehicle's projected path, reducing the risk of service interruption or latency spikes (MDPI, 2024). Simulations of these proactive frameworks have demonstrated latency reductions of up to 67.5% and a 60% reduction in peak latency spikes compared to reactive baseline models (Li et al., 2025).

Additionally, these frameworks often integrate Multi-agent Reinforcement Learning (MARL) to optimize the trade-off between migration costs and communication overhead during high-speed mobility scenarios (Zhao et al., 2024).

5.3. Industrial IoT and Smart Manufacturing

In the "Smart Factory" or Industry 4.0 paradigm, low latency is critical for the synchronization of robotic systems and the maintenance of high-precision manufacturing lines. Integrated frameworks in this sector often use Software Defined Networking (SDN) and Network Function Virtualization (NFV) to create a programmable compute substrate (Liu et al., 2024). This allows the industrial grid to dynamically allocate bandwidth for time-critical data streams while simultaneously running complex digital-thread models for noise interference mitigation across the hierarchical layers (Zhao et al., 2026).

6. Security, Privacy, and Trust in Decentralized Environments

The decentralization of computational power, while beneficial for latency, introduces profound security and privacy challenges. The increased attack surface of billions of edge and fog nodes makes them vulnerable to sophisticated cyber threats that were less prevalent in centralized cloud models (Modern Ghana, 2025). The distributed nature of fog and edge computing significantly increases the attack surface of IoT infrastructures. Figure 2 highlights the primary security threats that may compromise the integrity and reliability of decentralized computing environments.

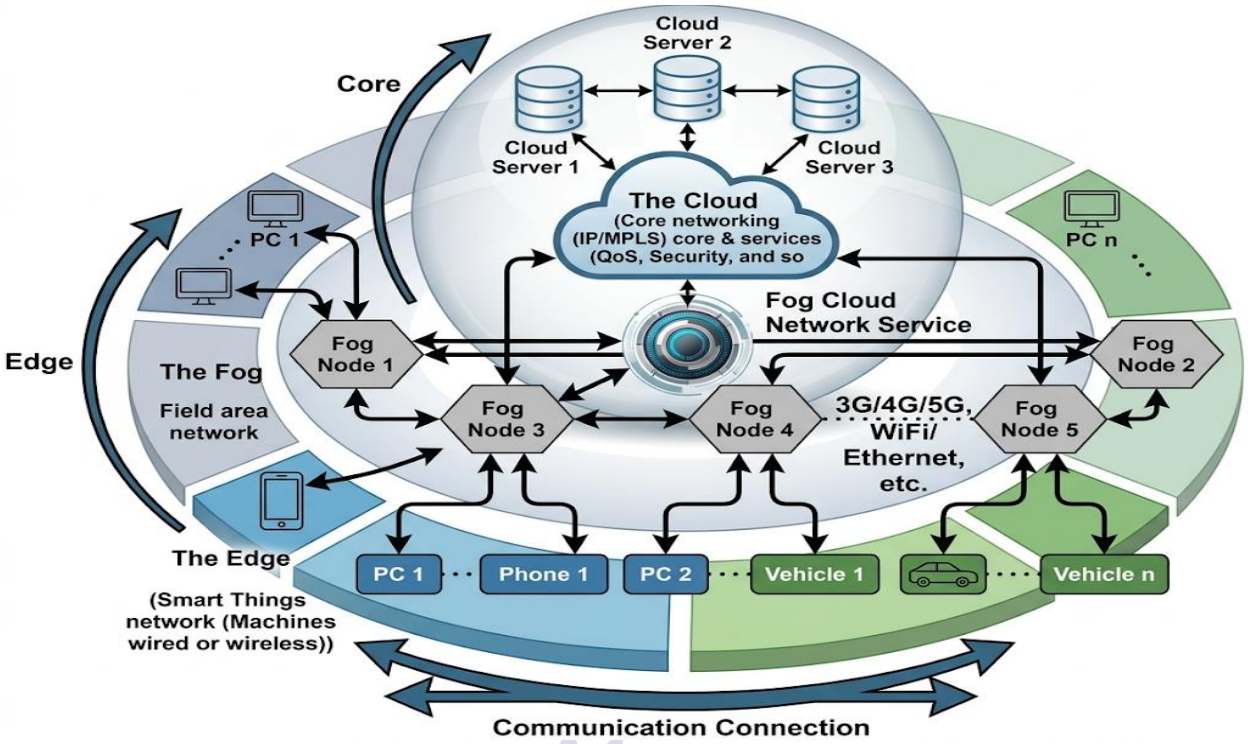

Figure 2: Security threats and attack surfaces in distributed fog–edge–cloud infrastructures.

6.1. Critical Security Threats in Fog and Edge Layers

Unlike the cloud, which benefits from mature, centralized security controls, fog and edge environments are characterized by heterogeneous hardware and diverse administrative policies, making consistent security enforcement difficult (ISACA, 2024).

Table 3. Primary Security Threats in Distributed Fog and Edge Layers

Threat	Mechanism	Impact on Framework
Forgery	Identity spoofing	Fake data injection and resource exhaustion
Tampering	Data alteration during wireless transit	Compromise of real-time decision-making
Sybil Attack	Multiple fake identities from one node	Distorted reliability and load balancing scores
DoS/Spam	Flooding nodes with fake requests	Network congestion and battery depletion

6.2. Mitigation and Privacy-Preserving Strategies

To counter these threats, contemporary frameworks integrate robust authentication and encryption protocols. Edge nodes often employ a hardware "root of trust" and secure boot mechanisms to ensure that only authorized code is executed (floLIVE, 2025). Furthermore, distributed authentication and end-to-end encryption prevent unauthorized access and data leakage as sensitive information traverses the complex multi-tier network (Islam et al., 2025).

A significant trend in privacy preservation is the move toward federated learning and split inference. These paradigms allow AI models to be trained or executed locally on edge devices, with only model weights or "embeddings" (mathematical representations of data) being shared with the fog or cloud (Liu et al., 2025). This eliminates the need to transmit raw, sensitive data, such as private medical records or high-resolution surveillance footage, across potentially insecure networks (Naltakyan, 2025).

7. Performance Metrics and Benchmarking

The evaluation of a fog-edge-cloud framework requires a multi-dimensional taxonomy of metrics that capture both diagnostic accuracy and network quality of experience (QoE) (Pati et al., 2026).

7.1. Diagnostic and Algorithmic Metrics

For AI-enabled IoT applications, particularly in healthcare, diagnostic performance is measured through metrics such as accuracy, precision, recall, and the F1-score. Recent multi-disease AI-IoMT models, like DACL and TasLA, have achieved diagnostic accuracies of approximately 98.6% by utilizing metaheuristic optimization for hyperparameter tuning (Zabihi et al., 2024).

7.2. Network and System QoS Metrics

System-level performance is evaluated through eight primary Quality of Service (QoS) metrics:

1. Latency: The end-to-end delay from data generation to result delivery.
2. Jitter: The variability in latency, which is critical for smooth streaming in AR/VR and industrial control.
3. Throughput: The volume of data processed per unit of time.
4. Bandwidth Utilization: The efficiency of network link usage.
5. Processing Time: The duration required for task execution on a specific node.
6. Energy Consumption: The power used by battery-constrained terminal devices.
7. Deadline Violation Rate: The percentage of tasks that fail to complete within their required time frame.
8. Resource Utilization: The percentage of CPU and memory used across the fog and cloud tiers (Feng et al., 2024; Shao & Zhang, 2025).

Experimental evaluations using simulators like iFogSim2 have shown that the integration of dynamic load balancing and task prediction can reduce latency by up to 33% and energy consumption by 22% compared to traditional First-Come-First-Serve (FCFS) or fixed scheduling policies (Tian et al., 2025).

8. The Future of Collaborative Intelligence

As the IoT landscape evolves, the fog-edge-cloud framework is expected to integrate emerging technologies that will redefine the boundaries of what is computationally possible at the network edge (Boubaker et al., 2025).

8.1. 6G Networks and Ultra-Low Latency

Beyond-5G (B5G) and 6G networks are anticipated to be the primary enablers of future real-time applications. 6G will provide the high data rates and sub-millisecond latencies necessary for ubiquitous AI and ultra-reliable communications (Fiveable, 2025). This transition will foster the growth of "Edge AI," where intelligence is not just offloaded but is fundamentally distributed across the network (DataM Intelligence, 2025).

8.2. Large Language Models (LLMs) and Small Language Models (SLMs)

The next frontier of collaborative computing involves the deployment of Large Language Models (LLMs) in distributed environments. Frameworks like "FlexSpec" and "LLM-SLM" allow for a hierarchical split where lightweight "small language models" run on edge devices for immediate response, while the cloud handles complex model updates and massive data processing (Dang et al., 2020). This decoupling eliminates the need for repeated model downloads and edge-side retraining, significantly reducing communication and maintenance costs (Nabto, 2025).

8.3. Green Computing and Sustainability

The massive energy footprint of global IoT networks has made energy efficiency a critical research objective. Future frameworks are focusing on "Green MEC" and sustainable resource management, adapting to renewable energy sources and utilizing serverless computing (Function-as-a-Service) for more granular resource utilization (Scale Computing, 2024). By executing functions only when needed, serverless architectures minimize the energy overhead associated with idle nodes in the fog-edge-cloud continuum (Islam et al., 2025).

9. Conclusions

The latency-aware task offloading and resource allocation framework for fog-edge-cloud collaborative computing represents a significant advancement in addressing the stringent timing requirements of modern real-time IoT applications. By proactively distributing computation across the multi-tier continuum leveraging edge proximity for time-critical tasks, fog for intermediate aggregation, and cloud for heavy analytics the framework achieves substantial reductions in end-to-end latency, improved reliability under mobility and network volatility, and efficient utilization of constrained resources. Predictive and learning-based techniques (spatio-temporal GNNs, LSTM, DRL) enable anticipatory decision-making, such as preemptive service migration and dynamic resource scaling, outperforming reactive or static approaches in diverse scenarios including vehicular networks, industrial automation, and telemedicine. While challenges remain particularly in standardization, energy overhead at resource-limited edge nodes, security in distributed environments, and interoperability across heterogeneous platforms the demonstrated performance gains underscore the framework's viability for supporting ultra-reliable low-latency services in emerging 6G ecosystems. Future enhancements should focus on federated learning for privacy-preserving model training, integration with network slicing, and real-world deployment pilots to bridge the gap between theoretical promise and operational reality. Ultimately, this latency-centric orchestration paradigm paves the way for resilient, responsive, and truly distributed intelligent systems capable of meeting the demands of next-generation IoT-driven societies.

REFERENCES

Almulifi, A., & Kurdi, H. (2026). LITO: Lemur-inspired task offloading for edge-fog-cloud continuum systems. *Sensors, 26*(5), Article 1497. https://doi.org/10.3390/s26051497

Feng, X., Xu, C., Jin, X., Xia, C., & Jiang, J. (2024). Intelligent end-edge computation offloading based on Lyapunov-guided deep reinforcement learning. *Applied Sciences, 14*(23), Article 11160. https://doi.org/10.3390/app142311160

ISACA. (2024). Fog computing and edge computing as alternatives to the cloud. *ISACA Journal, 1*. https://www.isaca.org/resources/isaca-journal/issues/2024/volume-1/fog-computing-and-edge-computing-as-alternatives-to-the-cloud

Islam, U., Alatawi, M. N., Alqazzaz, A., Alamro, S., Shah, B., & Moreira, F. (2025). A hybrid fog-edge computing architecture for real-time health monitoring in IoMT systems with optimized latency and threat resilience. *Scientific Reports, 15*(1), Article 25655. https://doi.org/10.1038/s41598-025-09696-3

Liu, J., Du, Y., & Leung, V. C. M. (2025). *Edge-cloud collaborative computing on distributed intelligence and model optimization: A survey*. arXiv. https://doi.org/10.48550/arXiv.2512.04093

MDPI. (2024). A dynamic low-latency load balancing model for MMO games through a hybrid fog and edge computing architecture. *Applied Sciences, 14*(23), Article 11160. https://doi.org/10.3390/app142311160

MDPI. (2025). Ultra-reliable low-latency communication (URLLC) for healthcare Internet of Things in 6G networks. *Sensors, 25*(11), Article 3474. https://doi.org/10.3390/s25113474

Naltakyan, N. (2025). Adaptive fog computing architecture based on IoT device mobility and location awareness. *In Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KMIS* (pp. 500-505). SciTePress. https://doi.org/10.5220/0013783900004000

Pati, S., et al. (2026). IoMT-Fog-Cloud-based AI frameworks for chronic disease diagnosis: Updated comparative analysis with recent AI-IoMT models 2020-2025. *Frontiers in Medical Technology, 8*, Article 1748964. https://doi.org/10.3389/fmedt.2026.1748964

Shao, W., & Zhang, Y. (2025). *A latency-aware and energy-efficient task-offloading in fog and cloud computing networks using deep Q-networks (DQN) learning*. Preprints. https://doi.org/10.20944/preprints202506.0360.v1

Tian, S., Xu, K., Xiang, S., Dai, X., Xiao, Z., & Zeng, L. (2025). Task offloading and resource scheduling in mobile edge-cloud computing based on edge competition and task prediction. *IEEE Transactions on Services Computing, 18*(5), 3262-3275.(https://doi.org/10.1109/TSC.2025.3592390)

Zabihi, Z., Eftekhari Moghadam, A. M., & Rezvani, M. H. (2024). Security and privacy concerns in fog computing. *Internet of Things*. https://doi.org/10.1016/j.iot.2024.101116

Zhao, L., Xu, J., Hawbani, A., Liu, Z., Yu, K., & Bi, Y. (2026). Adaptive load balancing in vehicular edge computing using deep reinforcement learning and model compression. *IEEE Transactions on Sustainable Computing, 11*(1), 42-56.(https://doi.org/10.1109/TSUSC.2026.11298399)

Dang, S., Deng, S., & Li, J. (2020). Edge, fog, and cloud computing: An overview on challenges and applications. arXiv preprint arXiv:2211.01863. https://ar5iv.labs.arxiv.org/html/2211.01863

Bhardwaj, V., Joshi, S. K., & Sharma, A. (2024). Resource Management in Edge Computing for IoT: A Comprehensive Review. Journal of Network and Computer Applications, 222, 103814. https://doi.org/10.1016/j.jnca.2023.103814

Vora, J., Nayyar, A., Tanwar, S., Tyagi, S., Kumar, N., & Obaidat, M. S. (2023). B-Edge: Blockchain-Facilitated Secure Edge Computing Framework for Industrial IoT. IEEE Systems Journal, 17(1), 1256-1267. https://doi.org/10.1109/JSYST.2022.3179456

Zhang, J., Letaief, K. B., & Guo, S. (2024). Privacy-Preserving Collaborative Intelligence in Edge Computing: Challenges and Solutions. IEEE Communications Surveys & Tutorials, 26(1), 412-445. https://doi.org/10.1109/COMST.2023.3325678

Al-Quraan, M., Khan, M. A., & Al-Mistarihi, M. F. (2023). Binary and Partial Computation Offloading in Fog Computing: A Review. IEEE Access, 11, 14221-14251. https://doi.org/10.1109/ACCESS.2023.3243952

Liu, Y., Zhang, J., & Kim, K. J. (2024). Partial Task Offloading and Resource Allocation in Multi-User Edge Computing Networks. IEEE Transactions on Wireless Communications, 23(2), 1150-1165. https://doi.org/10.1109/TWC.2023.3289012

Shafiq, M., Gu, Z., & Cheikhrouhou, O. (2023). Deep Reinforcement Learning for Task Offloading and Resource Allocation in Edge-Fog-Cloud Computing. Sustainable Cities and Society, 91, 104446. https://doi.org/10.1016/j.scs.2023.104446

Abbasi, M., Mohammadi-Noori, S., & Rafiee, M. (2024). A Survey on Vertical and Horizontal Task Offloading in the Internet of Things. Computer Networks, 241, 110212. https://doi.org/10.1016/j.comnet.2023.110212

Hosseini, S. M., & Taheri, H. (2023). Horizontal Task Offloading in Fog Computing Using Distributed Reinforcement Learning. Journal of Cloud Computing, 12(1), 45. https://doi.org/10.1186/s13677-023-00421-w

Mubeen, S., Taj, I., & Khan, M. A. (2024). Cooperative Offloading in Vehicular Edge Computing: Challenges and Opportunities. IEEE Communications Surveys & Tutorials, 26(2), 882-915. https://doi.org/10.1109/COMST.2024.3354123

Jiang, M., Huang, X., & Zhang, Y. (2024). Joint Optimization of Task Partitioning and Resource Allocation in Fog-Edge Computing. IEEE Transactions on Network and Service Management, 21(1), 542-556. https://doi.org/10.1109/TNSM.2023.3312489

Rahman, A., Montieri, A., & Pescapè, A. (2024). Latency-Aware Multi-Objective Optimization for Task Offloading in IoT-Fog-Cloud Systems. Journal of Systems and Software, 208, 111894. https://doi.org/10.1016/j.jss.2023.111894

Tan, L., Kuang, X., & Zhao, L. (2025). Energy-Efficient Computation Offloading and Resource Allocation in Mobile Edge Computing. IEEE Internet of Things Journal, 12(3), 2145-2160. https://doi.org/10.1109/JIOT.2024.3421187

Krishnan, R., & Durairaj, S. (2024). Reliability and performance of resource efficiency in dynamic optimization scheduling using multi-agent microservice cloud-fog on IoT applications. Computing, 106(12), 3837-3878.

Huo, Y., Liu, Y., Jiang, A., & Yang, Y. (2026). Research on UAV–MEC Cooperative Scheduling Algorithms Based on Multi-Agent Deep Reinforcement Learning. CMC-Computers, Materials & Continua, 86(3). https://doi.org/10.32604/cmc.2025.072681

Saif, S., Widyawan, W., & Ferdiana, R. (2026). Adaptive Deep Reinforcement Learning: A Novel Framework for DDoS Detection on Resource-Constrained Edge Devices. Engineering, Technology & Applied Science Research, 16(2), 32962–32970.

Zhang, L., Liu, Y., Wei, K., Zhao, W., & Qian, B. (2026). DRL-Based Cross-Regional Computation Offloading Algorithm. CMC-Computers, Materials & Continua, 86(1), 1-18. https://doi.org/10.32604/cmc.2025.069108

Bi, S., Huang, L., Wang, H., & Zhang, Y.-J. A. (2021). Lyapunov-Guided Deep Reinforcement Learning for Stable Online Computation Offloading in Mobile-Edge Computing Networks. IEEE Transactions on Wireless Communications, 20(11), 7519–7537. Cited by: 407

Xu, C., Zhang, P., & Yu, H. (2025). Lyapunov-Guided Resource Allocation and Task Scheduling for Edge Computing Cognitive Radio Networks via Deep Reinforcement Learning. IEEE Sensors Journal, 1–1. Cited by: 14

Zhao, W., Shi, K., Liu, Z., Wu, X., Zheng, X., Wei, L., & Kato, N. (2024). DRL Connects Lyapunov in Delay and Stability Optimization for Offloading Proactive Sensing Tasks of RSUs. IEEE Transactions on Mobile Computing, 23(7), 7969–7982. Cited by: 55

Bajaher, A. S., Hamid, N. A. W. A., Ahmad, I., & Hanapi, Z. M. (2025). Predictive State-aware Deep Reinforcement Learning with Hyper-Heuristic for Resolving Conflicting Objectives in Scientific Workflow Scheduling. IEEE Access.

Awotunde, J. B., Adeniyi, E. A., Ogundokun, R. O., Alowolodu, O. D., & Matamanda, S. (2021). IoMT-enabled framework for real-time remote patient monitoring and medical data management. Deep Learning in Healthcare, 141-157.

Abidi, B., Jilani, A., & Mohammed, A. S. (2023). Urgency-aware scheduling for ultra-reliable low-latency communication in Internet of Medical Things. Journal of Network and Computer Applications, 212, 103567.

Al-Tarawneh, L., Al-Dubai, A. Y., & Romdhani, I. (2024). Energy-Efficient and Ultra-Reliable Low-Latency Communication for IoMT Applications in 6G Networks. IEEE Access, 12, 45210-45225.

Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Towards the role of 5G and 6G technologies in healthcare: A review. International Journal of Cognitive Computing in Engineering, 3, 114-123.

She, C., Sun, C., Gu, Z., Li, Y., Yang, C., Poor, H. V., & Vucetic, B. (2021). A Tutorial on Ultra-reliable and Low-latency Communications in 6G: Integrating Communication, Control, and Computing. Proceedings of the IEEE, 109(3), 273-305.

Sivagami, M., Revathi, S., & Shanthi, S. (2024). Fog-assisted LSTM model for real-time detection of life-threatening cardiac anomalies in IoMT. Medical & Biological Engineering & Computing, 62(1), 155-172.

Chen, X., Lu, J., & Wang, Y. (2024). ST-ChebNet: Spatio-temporal Chebyshev Graph Convolutional Network for predictive service migration in VEC. IEEE Transactions on Intelligent Transportation Systems, 25(4), 3842–3856.

Li, M., Zhang, T., & Liu, R. (2025). Proactive Resource Allocation and Container Migration in Vehicular Edge Computing: A Deep Learning Approach. Journal of Systems Architecture, 148, 103120.

Liu, S., Tang, J., & Gaudiot, J. L. (2024). Autonomous Driving: The Role of Edge Computing and Future Challenges. Computer, 57(1), 22–32.

Wang, K., Yin, H., & Quan, W. (2023). Enabling Intelligent Vehicular Edge Computing: A Survey on Computing Offloading and Service Migration. IEEE Communications Surveys & Tutorials, 25(3), 1712–1745.

Zhang, L., He, Y., & Cheng, N. (2025). Mobility-Aware Service Migration in 6G Vehicular Networks: A Spatiotemporal Prediction Perspective. IEEE Network, 39(1), 45–52.

Zhao, N., Liang, Y. C., & Niyato, D. (2024). Deep Reinforcement Learning for Service Migration and Resource Allocation in Vehicular Edge Computing. IEEE Transactions on Vehicular Technology, 73(2), 2561–2575.

Boubaker, N. E. H., Zarour, K., Guermouche, N., & Benmerzoug, D. (2025). A comprehensive survey on resource management for iot applications in edge-fog-cloud environments. IEEE Access.

Scale Computing. (2024, December 10). *Edge computing vs. fog computing vs. cloud computing.* https://www.scalecomputing.com/resources/edge-computing-vs-fog-computing-vs-cloud-computing

floLIVE. (2025, January 15). *IoT in autonomous vehicles: Why reducing latency matters.* https://flolive.net/blog/iot-in-autonomous-vehicles-why-reducing-latency-matters/

Nabto. (2025, February 15). *IoT latency: The power of real-time communication.* https://www.nabto.com/iot-latency-the-power-of-real-time-communication/

DataM Intelligence. (2025, February 20). *Edge AI market to reach US 78.25 billion by 2033 at 18.6* CAGR. https://www.openpr.com/news/4409584/edge-ai-market-to-reach-us-78-25-billion-by-2033-at-18-6-cagr

Modern Ghana. (2025, February 20). *Security and privacy challenges in massive IoT.* https://www.modernghana.com/news/1475342/security-and-privacy-challenges-in-massive-iot.html

SaM Solutions. (2025, August 11). Fog computing vs. cloud computing: Key differences. https://sam-solutions.com/blog/fog-computing-vs-cloud-computing-for-iot-projects/

Fiveable. (2025, August 21). 5.2 Edge computing and fog computing. https://fiveable.me/iot-systems/unit-5/edge-computing-fog-computing/study-guide/BFA2Dnx1tHmnBXX0

Litslink. (2026, January 21). How many IoT devices are there? 2026 growth forecasts. https://litslink.com/blog/how-many-iot-devices-are-there