International Journal of Multidisciplinary Research and Development
www.allsubjectjournal.com
ISSN Online: 2349-4182
Received: 01-06-2026, Accepted: 01-07-2026, Published: 16-07-2026
Volume 13, Issue 3, 2026, Page No. 46-49

Adaptive AI scheduling across edge-cloud continuum

Dr. G Sripriya[1], Darun M K[2], Nandana K[2]
[1]Assistant Professor, Department of CA, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India
[2] Department of CA, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India
DOI: https://doi.org/10.66856/ijmrd.2026.13.3.13329

Abstract
The Edge–Cloud Continuum integrates edge nodes, fog layers, and cloud infrastructures to support intelligent, low-latency AI workloads, yet existing scheduling approaches suffer from static policies and single-objective optimisation. To address this, an Adaptive AI-driven Task Scheduling (AATS) framework is proposed, combining Deep Q-Network reinforcement learning with deadline-safety heuristics and federated model updates across continuum tiers. Results demonstrate a 65.4% latency reduction, 54.8% lower energy consumption, and 89.5% higher throughput over static baselines, confirming the hybrid Edge–Cloud Continuum as the optimal paradigm for next-generation AI systems.

Keywords: Edge computing, cloud computing, edge–cloud continuum, artificial intelligence, federated learning, task offloading, resource orchestration, 6G networks, AATS framework

Introduction

Cloud computing has long served as the backbone of large-scale AI workloads, but applications such as autonomous vehicles, smart healthcare, and industrial automation demand real-time responsiveness
that centralised architectures cannot deliver resulting in increased latency, bandwidth consumption, and privacy concerns [1]. To address these limitations, the Edge–Cloud Continuum integrates IoT devices, edge nodes, fog layers, and cloud data centres into a unified environment where workloads are intelligently distributed across tiers according to latency, resource availability, and application priorities [1, 5], Artificial Intelligence has become the key enabling technology for managing these distributed environments, allowing systems to predict workload changes, optimise resource allocation, schedule tasks autonomously, and enable real-time decision-making at the edge [2, 3], This paper reviews recent advances spanning continuum architectures, AI-driven scheduling, federated learning, hardware-accelerated orchestration, and cognitive AI deployment. It further identifies critical scheduling limitations in existing systems and proposes the Adaptive AI-driven Task Scheduling (AATS) framework as a novel solution for next-generation AI workload management across the continuum [1–15]

Literature review

Rosendo *et al*. consolidated methods for enabling machine learning and deep learning inference, centralised and distributed training, and big data stream analytics across the Edge Cloud Continuum. Their systematic review surveyed federated learning, online learning, and transfer learning paradigms and evaluated experimental testbeds for reproducibility, identifying the absence of a holistic performance model as a critical open challenge [1].

Liu *et al*. surveyed edge-cloud collaborative computing targeting distributed intelligence and model optimisation, documenting exponential citation growth in edge-cloud collaboration research peaking in 2024. Three principal contribution categories were identified: model optimisation for resource constrained deployment, collaborative inference and split computing, and privacy-preserving distributed training [2].

Naeem *et al*. conducted a forward-looking survey of federated learning for edge AI, analysing over 200 papers and organising FL designs into cross-device, cross-silo, edge-only, decentralised, and hybrid categories. Edge AI and federated learning were established as complementary technologies, with key unresolved concerns including robustness, sustainability, and single-point-of-failure risks [3].

Naeem *et al*. proposed an AI-powered task scheduling system combining the Unfair
Semi-Greedy (USG) algorithm, Earliest Deadline First (EDF), and Enhanced Deadline Zero-Laxity (EDZL) with reinforcement learning adaptive logic. A dynamic resource table selects the best scheduler based on current load and task criticality, outperforming single-algorithm baselines in deadline adherence and resource utilisation [4].

Belcastro *et al*. mapped the current engineering landscape of the Edge–Cloud Continuum by covering static and dynamic task offloading strategies, the multi-tier client-server model, and microservice placement strategies. Quantitative findings demonstrated that offloading computational tasks to far-edge and near-edge servers significantly reduces latency and energy consumption compared to direct cloud access [5].

Sannapureddy *et al*. surveyed architectural models and scheduling strategies for applications with stringent latency budgets, demonstrating that effective orchestration requires jointly modelling network conditions, workload characteristics, and end-to-end latency budgets. Static offloading policies were identified as inadequate for dynamic real-world workloads, necessitating context-aware dynamic decision-making [6].

Blanquez *et al*. proposed a multi-tier orchestration framework targeting the low-latency and high-bandwidth demands of 6G networks, dynamically provisioning hardware acceleration through FPGAs, GPUs, and NPUs based on AI workload profiles. The framework integrates with Multi-access Edge Computing deployments and

demonstrates up to three times throughput improvement over software-only orchestration [7].

Ofrim *et al*. addressed the integration of IoT devices with smart-grid edge-fog-cloud architectures, employing offloading strategies using metaheuristics for near-optimal solutions and reinforcement learning for adaptive control. Containerised workloads managed via Kubernetes-based orchestration demonstrated a 40 percent improvement in service availability over single-tier deployments [8].

Wang *et al*. reviewed federated continual learning (FCL), which fuses knowledge across distributed edge clients while preserving data privacy and retaining prior knowledge. FCL directly addresses catastrophic forgetting in non-stationary distributed environments a critical limitation for long-running edge AI deployments such as autonomous vehicles and industrial IoT [9].

Mishra *et al*. conducted a systematic review of federated learning within the edge-cloud continuum, analysing communication efficiency strategies including gradient compression, asynchronous aggregation, and hierarchical FL. A central finding confirmed that hierarchical FL where local aggregation at edge servers precedes global aggregation in the cloud offers the best balance between communication cost and model convergence speed [10].

Sofia *et al*. documented the transition from the Edge–Cloud Continuum to the Cognitive Computing Continuum (CCC), presenting the ENACT framework as a reference architecture integrating IoT devices, edge servers, and cloud HPC resources under a unified cognitive orchestration plane. Trust, explainability, and regulatory compliance with the EU AI Act were identified as principal open challenges [11].

Casamayor Pujol *et al*. proposed using active inference grounded in the Free Energy Principle from computational neuroscience for distributed decision-making across the continuum. Each node maintains a generative model of its environment and minimises variational free energy to optimise resource allocation, offering a biologically inspired alternative to reinforcement learning with faster convergence than deep RL baselines [12].

Semerikov *et al*. surveyed techniques for deploying large language models at the edge, including quantisation, pruning, knowledge distillation, speculative decoding, and split inference. The OLRAMT-DEC framework dynamically partitions LLM inference across device, edge, and cloud tiers, achieving a 55 percent reduction in first-token latency versus full cloud inference while maintaining comparable accuracy [13].

Thota identified that cloud-designed AI models are overprovisioned for edge inference tasks, resulting in wasted compute and excess latency. A hierarchical edge architecture spanning IoT devices, edge nodes, gateways, and cloud backends demonstrated latency reductions of up to 60 percent versus cloud-only deployment through early-stage data filtering at edge nodes [14].

Sathupadi *et al*. investigated resource allocation strategies for AI-driven predictive maintenance tasks, formulating a dual-objective optimisation balancing energy consumption and bandwidth usage while preserving low-latency operation. Experiments on industrial sensor networks demonstrated that edge-cloud synergy reduces bandwidth consumption by 45 percent and operating energy by 30 percent compared to cloud-only baselines [15].

Problem statement

Existing task scheduling approaches in the Edge–Cloud Continuum suffer from three critical limitations. First, static offloading policies fail to adapt to dynamic workload changes, fluctuating network conditions, and heterogeneous resource availability across tiers resulting in latency violations and resource underutilisation [4, 6]

Second, single-algorithm schedulers such as Earliest Deadline First (EDF) or Round Robin cannot simultaneously optimise for latency, energy, and throughput across multi-tier environments. Third, current frameworks lack a unified intelligence layer capable of making real-time, context-aware scheduling decisions that span IoT devices, edge nodes, fog layers, and cloud data centres simultaneously [1, 5], These limitations result in degraded Quality of Service (QoS), increased energy consumption, and missed deadlines in latency-sensitive AI applications such as autonomous vehicles, smart healthcare monitoring, and industrial predictive maintenance. There is therefore a critical need for an AI-driven, adaptive, multi-tier task scheduling framework capable of dynamically optimising resource allocation decisions across the full Edge–Cloud Continuum

Existing systems

Current task scheduling systems in the Edge–Cloud Continuum can be categorised into three principal approaches based on their decision-making mechanisms and optimisation objectives

1. Static offloading policies

Traditional static offloading policies assign tasks to fixed tiers based on predefined rules such as task size thresholds or application type. While computationally simple, these approaches cannot respond to dynamic changes in network latency, node availability, or workload burst patterns. Studies confirm that static policies result in latency overruns of up to 40 percent under peak load conditions compared to dynamic alternatives [5, 6],

2. Single algorithm schedulers

Conventional schedulers such as Earliest Deadline First (EDF), Round Robin (RR), and First Come First Served (FCFS)

optimise for a single objective typically deadline adherence or fairness without considering energy consumption, throughput, or cross-tier resource availability. These approaches perform acceptably in homogeneous cloud environments but degrade significantly in heterogeneous multi-tier continuum deployments [4].

3. Deep reinforcement learning schedulers

Recent works have introduced Deep Reinforcement Learning (DRL) schedulers that learn optimal offloading policies through environment interaction. While these approaches outperform static and single-algorithm methods, they suffer from slow convergence, high training overhead, and instability under rapidly changing network conditions limiting their applicability for real-time edge AI workloads [4, 12],

Table 1: Limitations of Existing Scheduling Approaches

APPROACH	LATENCY	ENERGY Eff.	ADAPTABILITY	MULTI-TIER
Static Offloading	Poor	Poor	None	No
EDF/Round Robin	Moderate	Poor	Low	No
Deep RL Schedule	Good	Moderate	High	Partial
AATS (Proposed)	Excellent	Excellent	High	Yes

Proposed framework-AATS

This paper proposes the Adaptive AI-driven Task Scheduling (AATS) framework a multi-tier, context-aware scheduling system that dynamically distributes AI workloads across IoT devices, edge nodes, fog layers, and cloud data centres based on real-time system state. AATS addresses all three limitations identified through a hybrid intelligent design. The AATS framework consists of three functional layers operating in continuous feedback.

Monitoring Layer - The Monitoring Layer continuously collects real-time metrics including CPU utilisation, memory availability, network latency, bandwidth, task queue length, and battery level from all continuum nodes. A lightweight telemetry agent runs at each tier, forwarding state vectors to the Decision Engine every 100 ms.

Decision Engine - The Decision Engine is the core intelligence component employing a hybrid scheduling policy that combines: (i) a Deep Q-Network (DQN) agent trained to select the optimal tier for each incoming task based on the current state vector; and (ii) a heuristic override module applying Enhanced Deadline Zero-Laxity (EDZL) logic when task deadlines are within a critical threshold. This hybrid design ensures both long-term optimality and immediate deadline safety [4, 12], Execution Layer - The Execution Layer receives scheduling decisions and deploys tasks via containerised microservices across the selected tier. A feedback loop returns execution outcomes actual latency, energy consumed, and deadline met or missed to the DQN agent for continuous online learning [8].

The scheduling decision at each timestep t is formulated as an optimal policy derived from the action-value function:

$$\pi(s\_t) = \text{argmax}\_a\, Q(s\_t, a; \theta)^*$$

where s_t is the state vector comprising [CPU_util, Mem_avail, Net_latency, Queue_len, Task_deadline, Task_size], a $\in$ {IoT, Edge, Fog, Cloud} is the tier assignment action, Q(s_t, a; θ) is the action-value function approximated by a three-layer neural network with parameters θ, and π^* is the optimal scheduling policy.

The reward function balances latency, energy, and deadline compliance:

$$R\_t = \alpha(1 - L\_norm) + \beta(1 - E\_norm) - \gamma \cdot D\_miss$$

where L_norm is normalised latency, E_norm is normalised energy consumption, D_miss is a binary penalty for deadline violation, and $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ are weighting coefficients prioritising latency while penalising deadline misses.

To ensure the DQN agent adapts across distributed deployments without centralising raw telemetry data, AATS incorporates a federated model update mechanism. Each edge node maintains a local DQN replica. Every 500 scheduling cycles, local model gradients are aggregated at the fog layer using the FedAvg algorithm, and the updated global model is redistributed preserving data privacy while enabling continuum-wide collaborative learning [9, 10].

This federated design ensures AATS operates effectively across geographically distributed deployments with heterogeneous network conditions, without requiring centralised data collection or single-point orchestration control.

Experimental results

The AATS framework was evaluated using a simulated Edge–Cloud Continuum environment implemented in Python 3.11, comprising 50 IoT sensor nodes, 10 edge servers (4-core, 8GB RAM), 3 fog aggregation nodes (16-core, 32GB RAM), and 1 cloud backend (128-core, 512GB RAM). Network latency was modelled as: IoT→Edge: 5–15ms, Edge→Fog: 15–40ms, Fog→Cloud: 40–120ms. A workload generator produced 10,000 heterogeneous tasks with deadlines ranging from 50ms to 2,000ms and sizes from 1MB to 500MB, following a Poisson arrival distribution ($\lambda = 20$ tasks/second). AATS was compared against four baseline algorithms: Static Offloading (SO), Earliest Deadline First (EDF), Round Robin (RR), and standalone Deep Q-Network (DQN-only).

Table 2: Comparative of Performance Results

ALGORITHM	AVG.LATENCY (ms)	DEADLINE MISS(%)	ENERGY (j/task)	THROUGHPUT (tasks/s)
Static Offloading	284.3	34.2	8.71	11.4
Round Robin	231.7	28.6	7.93	13.8
EDF	198.4	19.3	7.45	15.2
DQN-Only	142.6	11.7	5.82	17.9
AATS (PROPOSED)	98.3	4.1	3.94	21.6

AATS achieves an average latency of 98.3ms a 65.4% reduction versus Static Offloading and a 31.1% reduction versus DQN-Only. The deadline miss rate of 4.1% is the lowest across all evaluated algorithms, confirming the effectiveness of the EDZL heuristic override for critical tasks.

Energy consumption per task is reduced by 54.8% compared to Static Offloading, validating the energy-aware reward function design. Throughput improves by 89.5% over the Static Offloading baseline and 20.7% over DQN-Only, demonstrating the benefit of hybrid intelligence over pure learning approaches.

The AATS DQN agent converges to a stable scheduling policy within 800 training episodes, compared to 1,400 episodes for the standalone DQN-only baseline a 42.9% faster convergence attributable to the EDZL heuristic override reducing unnecessary exploration in critical deadline scenarios.

Federated model updates further improve convergence stability across distributed edge nodes, with cross-node policy variance reducing by 38% after three federated aggregation rounds. This confirms that AATS scales effectively across distributed continuum deployments without centralised data collection.

Conclusion

The Edge–Cloud Continuum has emerged as the most promising distributed computing paradigm for modern AI systems, overcoming the fundamental limitations of pure edge and cloud deployments by integrating IoT devices, edge nodes, fog layers, and cloud data centres into a unified heterogeneous environment [1, 5],

This paper reviewed fifteen recent studies across five thematic clusters systematic surveys, federated learning, workload scheduling, hardware orchestration, and cognitive AI and identified static and single-objective scheduling as the principal performance bottleneck. To address this, the Adaptive AI-driven Task Scheduling (AATS) framework was proposed, combining Deep Q-Network reinforcement learning with EDZL deadline-safety heuristics and federated model updates across continuum tiers.

Experimental evaluation against four baseline algorithms demonstrates that AATS achieves a 65.4% reduction in average latency, a 4.1% deadline miss rate, 54.8% lower energy consumption per task, and 89.5% higher throughput compared to static offloading confirming that intelligent, adaptive, multi-tier scheduling is the key to unlocking the full potential of the Edge–Cloud Continuum for AI systems [4, 7, 9, 12]., The comparative verdict is clear: edge nodes are optimal for real-time inference and privacy-sensitive applications; fog layers are best suited for hierarchical federated aggregation; cloud resources remain indispensable for large-scale training; and hybrid continuum architectures orchestrated with AI-driven scheduling represent the optimal solution for next-generation AI systems [4, 7, 9, 12] Future work will extend AATS to incorporate 6G network slicing, carbon-aware scheduling, and adversarial-robust federated aggregation for mission-critical deployments across smart cities, Industry 5.0, and autonomous transportation networks [7, 9, 15].

Reference

1. Rosendo D, Costan A, Valduriez P, Antoniu G. "Distributed Intelligence on the Edge-to-Cloud Continuum: A Systematic Literature Review," Journal of Parallel and Distributed Computing,2024:166:71–94.
2. Liu J, Du Y, Yang K, Wu J, Wang Y, Hu X, *et al*. "Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey," arXiv preprint arXiv:2505.01821, 2025.
3. Naeem AB, Senapati B, Rasheed J, *et al*. "A Comprehensive Survey of Federated Learning for Edge AI: Recent Trends and Future Directions," Preprints.org, 2025.
4. Naeem AB, Senapati B, Rasheed J, *et al*. "An Intelligent Job Scheduling and Real-Time Resource Optimization for Edge-Cloud Continuum in Next Generation Networks," Scientific Reports,2025:15(41534).
5. Belcastro L, Marozzo F, Orsino A, Talia D, Trunfio P. "Navigating the Edge-Cloud Continuum: A State-of-Practice Survey," arXiv preprint arXiv:2506.02003, 2025.
6. Anonymous *et al*. "Edge–Cloud Continuums for Latency-Sensitive Tasks," International Journal of AI, BigData, Computational and Management Studies (IJAIBDCMS), 2024.
7. Blanquez AL, *et al*. "Hardware-Accelerated Edge AI Orchestration on the Multi-Tier Edge-to-Cloud Continuum," Journal of Network and Systems Management, Springer, 2025.
8. Anonymous *et al*. "Cloud-Edge Continuum Orchestration for Containerized Workloads," IEEE, 2025.
9. Wang Z, Wu F, Yu F, Zhou Y, Hu J, Min G. "Federated Continual Learning for Edge-AI: A Comprehensive Survey," ACM Computing Surveys, arXiv preprint arXiv:2411.13740, 2024.
10. Mishra SK, Sahoo SK, Swain CK. "A Systematic Review on Federated Learning in Edge-Cloud Continuum," SN Computer Science,2024:5(7):887.
11. Sofia RC, *et al*. "Cognitive Computing Continuum: State-of-the-Art Review and ENACT Vision and Approach," Journal of Grid Computing, Springer, 2025.
12. Casamayor Pujol V, Sedlak B, Salvatori T, Friston K, Dustdar S. "Distributed Intelligence in the Computing Continuum with Active Inference," arXiv preprint arXiv:2505.24618, 2025.
13. Anonymous *et al*. "Edge Intelligence Unleashed: A Survey on Deploying Large Language Models in Resource-Constrained Environments," Journal of Edge Computing, 2025.
14. Anonymous *et al*. "Optimizing Edge Computing and AI for Low-Latency Cloud Workloads," International Journal of Science and Research Archive,2024:13(01):3484–3500.
15. Anonymous *et al*. "Edge-Cloud Synergy for AI-Enhanced Sensor Network Data: A Real-Time Predictive Maintenance Framework," Sensors, MDPI,2024:24(24):7918.