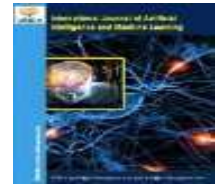


SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>

Research Paper

Open Access

An Adaptive Edge-Cloud Computing Architecture for Real-Time Intelligent Resource Allocation Using Machine Learning

Dr. Katakam Srinivasa Rao¹, T. Sunitha², M. Srirama Lakshmi Reddy³, Dr. Deepjyoti Kalita⁴, Shalini. S⁵, Dr Vikash kumar⁶¹CMR INSTITUTE OF TECHNOLOGY, HYDERABAD, Telangana -501401, Specialization in COMPUTER SCIENCE AND ENGINEERING, E-mail: srinivas.katakam@cmritonline.ac.in, ORCID ID: 0009-0007-9092-6366,²Assistant Professor, Hyderabad Institute of Technology and Management, HYDERABAD, Telangana -501401, Specialization in Computer Science and Engineering (AI & ML), E-mail: sunitha.cse@hitam.org, ORCID ID: 0009-0001-0917-2517³Associate Professor, CMR INSTITUTE OF TECHNOLOGY, HYDERABAD, Telangana -501401, Specialization in COMPUTER SCIENCE AND ENGINEERING, E-mail: lakshmireddy.msr@cmritonline.ac.in, ORCID ID: 0000-0001-8669-0812,⁴Assistant Professor, Department of Computer Science & IT, Mangaldai College (Autonomous), Darrang, E-mail: deepjyoti111@gmail.com, ORCID ID: 0000-0002-8205-5464⁵Associate Professor, Dayananda Sagar Academy of Technology and Management, E-mail: shalini.siddamallappa@gmail.com, ORCID:0000-0003-3809-4252⁶Assistant Professor, Department of Electrical Engineering, Ramchandra Chandravansi University, Palamu, Jharkhand. E-mail: Vikash9006@gmail.com

ABSTRACT

The convergence of edge and cloud computing has created a flexible computational continuum for supporting latency-sensitive, data-intensive, and intelligent applications. However, dynamic workloads, heterogeneous resources, mobility, and changing network conditions make real-time resource allocation increasingly complex. This review examines adaptive edge-cloud computing architectures for intelligent resource allocation using machine learning (ML), emphasizing architectural design, resource management, learning strategies, performance optimization, applications, and emerging research directions. Supervised, unsupervised, deep, reinforcement, multi-agent, federated, and graph-based learning approaches are discussed in relation to workload prediction, task offloading, resource provisioning, service migration, and distributed decision-making. The review highlights the importance of closed-loop adaptation, where resource states are continuously monitored, future demands are predicted, allocation decisions are executed, and resulting performance is evaluated for subsequent policy refinement. Key optimization objectives include latency, energy efficiency, resource utilization, throughput, cost, quality of service, and reliability. Applications across IoT, industrial systems, smart cities, connected vehicles, healthcare, and next-generation networks demonstrate the broad relevance of adaptive resource intelligence. Persistent challenges include heterogeneity, scalability, communication overhead, security, privacy, model drift, and generalization. Future research should prioritize continual learning, explainable intelligence, digital twins, green computing, AI-native 6G, and autonomous orchestration for scalable, secure, sustainable, and resilient real-time edge-cloud computing environments worldwide.

Keywords: Edge-Cloud Computing; Machine Learning; Resource Allocation; Adaptive Resource Management; Real-Time Computing

1. Introduction

The Internet of Things (IoT) is characterized by the recent surge in the number of connected devices, intelligent sensors, cyber-physical systems, and applications with artificial intelligence (AI) capabilities that are fundamentally changing the computation needs of modern digital infrastructures. While conventional cloud computing offers a huge amount of processing and storage capacity, it has one major limitation: it uses centralized data centers, which cause communication latency, network congestion, bandwidth usage and service disruptions when used by geographically distributed and latency-sensitive applications. Edge computing is a complementary paradigm that brought computation, storage, and analytics closer to the data sources and end users. But the demands for computation at the edge nodes are continuously growing and they are limited in resources. As a result, edge and cloud architectures are becoming more connected as edge-cloud continuum, allowing workloads to be flexibly scheduled among the various computing tiers. This integration brings the edge resources' low-latency processing capabilities and the computational scalability of cloud platforms together to establish a solid basis for responsive and scalable intelligent services (Hossain et al., 2023).

Efficient resource allocation is a core element of this continuum as the amount of computation, network quality, and availability of devices and application priorities can change significantly over time. Real-time applications are a special case where the constraints are very high, as allocation must be done quickly, and the application must meet latency requirements, along with energy, throughput, reliability, and QoS requirements. Traditional allocation mechanisms are often based on fixed rules, mathematical algorithms, static thresholds, or optimization routines based on relatively fixed assumptions. While they are all applicable in controlled

scenarios, they don't work so well in stochastic workloads, when the infrastructure changes, or when the dimensionality of the problem increases. Hence, scheduling mechanisms that can adapt to such temporal changes in workload and system state are desired for dynamic edge-cloud environments, in addition to those that are predetermined by policies (Tuli et al., 2020). This problem is more complicated if the computational, communication, storage, and energy resources have to be coordinated together across different devices and tiers of infrastructure.

Because of these limitations, machine learning (ML) offers a valuable way of addressing these challenges through the ability to better learn from past and current operating data, and to make resource-management decisions that reflect that learning. The workload forecasting, resource-demand prediction, task classification, congestion estimation, service placement, computation offloading and dynamic scheduling can be supported by ML models. Deep learning also allows for modelling of complex nonlinear relationships between workload properties and infrastructure states, while reinforcement learning allows resource controllers to adapt their allocation policies through multiple interactions with the dynamic environments. As a consequence of the increasing relevance of these techniques, machine and deep learning have been widely employed in the context of resource allocation in multi-access edge computing, with intelligent techniques being studied for the optimization of heterogeneous and conflicting objectives of a system (Djigal et al., 2022). A sequential decision-making problem in which allocation actions affect both current system performance and future operating states is particularly promising through the use of deep reinforcement learning (Wang et al., 2019).

The shift from isolated optimization algorithms based on ML to adaptive edge-cloud architectures is thus of particular importance. An efficient intelligent architecture should not only create an optimized allocation at a single moment, but it should also continuously monitor the infrastructure state, anticipate future resource needs, make context-aware allocation decisions, consider the impacts and make adjustments to future policies. These recent cloud-edge hybrid deep-learning frameworks show how distributed computational infrastructure can be combined with intelligent optimisation to enhance the scalability and utilization of resources in IoT systems. Recent cloud-edge hybrid deep-learning frameworks illustrate how distributed computational infrastructure may be coupled with intelligent optimisation to boost both scalability and resource appropriation in IoT systems. These architecture can enable workload migration from edge to cloud tier, dynamically provision resources to meet the needs of the workload, redirect workloads during congestion, and adjust workload priorities. Adaptive intelligence is particularly crucial for real-time data streams, where the distribution and nature of the data and the operating conditions constantly change, leading to the possible inaccuracy of the models and allocations created from static data sources (Sicakyüz et al., 2024).

Though considerable advances have been made, the current literature is divided into the fields of edge computing, cloud computing, task scheduling, computation offloading, prediction using ML, deep reinforcement learning, networking, and application-specific optimisation. Research often focuses on a single resource category or a specific component of a single performance goal, instead of the entire adaptive decision-making process through the edge-cloud continuum. Furthermore, since the architectures are different, the workloads differ, the optimizations are different, and the environments and performance metrics differ from one approach to the next, it can be difficult to compare performance directly across the different approaches. These restrictions define the necessity to consider the integration of architectural design and learning-driven resource intelligence and the critical investigation of adaptive mechanisms' ability to address operating conditions with diversity and change.

Therefore, the convergence of edge-cloud computing and machine learning in adaptive real-time resource allocation is critically analysed in this review. It brings together concepts from architecture, resource allocation, significant ML paradigms, adaptive resource management, and the optimization of performance goals, and investigates them for various emerging intelligent systems. Special focus is placed on comprehending ways in which learning processes can make the process of resource management from a state of "static" or "reactive" resource allocation to "predictive" and "context-aware" and "continually adaptive" orchestration. The review also investigates representative applications and evaluation environments, pinpoints methodological and deployment challenges, and highlights the new research opportunities of distributed intelligence, explainability, energy efficiency, security, continuous learning and autonomous orchestration. In this integrated view, the review provides a conceptual framework for designing novel edge-cloud resource allocation systems that make timely intelligent and adaptive allocations of resources. The specific objectives of this review are:

- To examine ML-driven mechanisms for adaptive and real-time resource allocation across the edge-cloud computing continuum.
- To evaluate existing resource-allocation approaches based on latency, energy efficiency, scalability, resource utilization, cost, and QoS.
- To identify key challenges, research gaps, and future directions for intelligent and autonomous edge-cloud resource management.

2. Edge-Cloud Architecture

Cloud, fog and edge computing are complementary paradigms of distributed computing that vary mainly by

their proximity to data sources, their computational power, and their response properties. Cloud computing has centralized, scalable processing and storage though it may bring in transmission latency and bandwidth overhead. Cloud and fog computing are similar in that they both rely on processing located at intermediate points in the network; however, fog computing processes data at the fog node, while edge computing processes data at a closer location to the source of data generation to ensure more immediate response time and less data transmission. These paradigms enable hierarchical resource allocation, where workloads are allocated based on computational requirements, latency, and resource availability (Shanmugam et al., 2023). The edge-cloud continuum brings end devices, edge nodes, intermediate infrastructure and centralized clouds together into a single resource continuum. These tiers can be partitioned or migrated based on network conditions, application urgency and capacity. This is especially useful for ML applications as it enables lightweight inference to take place at the edge while more powerful training and analytics are conducted in the cloud. With coordinated edge-cloud infrastructures, the execution of ML workloads can then be enhanced by leveraging resources from various computing tiers (Almurshed et al., 2025).

The typical layers of an adaptive edge-cloud architecture are the device layer, edge layer, communication layer, cloud layer, and orchestration layer. Data generation is done by end devices; localized processing and storage is provided by edge nodes; communication networks are used to connect distributed resources; and cloud infrastructure offers high capacity computation and persistent storage. Monitoring, scheduling, provisioning, workload placement and migration are orchestrated by an intelligent orchestration layer. By further hiding the complexity of the infrastructure, intelligent middleware can also enable the management of heterogeneous edge-cloud resources (Shah, 2021). Edge-Cloud infrastructure architectures are currently facing another big challenge – resource heterogeneity – as it is known that processors, GPUs, accelerators, memory, storage and communication links have varying capabilities. The computational load, latency and energy consumption are also widely different for applications (Figure 1). Therefore, resource allocation needs to be based on several characteristics and not just CPU capacity. Scheduling in heterogeneous end-edge-cloud solutions can be assisted by learning-based scheduling to achieve better workload mapping to appropriate computing resources (Shen et al., 2025).

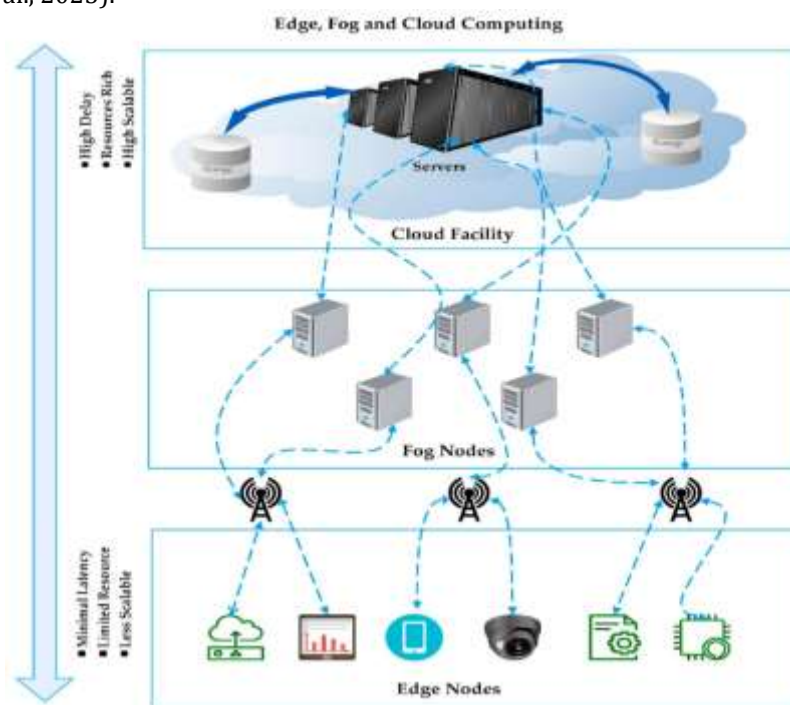


Figure 1. Edge-fog-cloud computing architecture (Khan et al., 2025).

Real-time applications also have rigid requirements such as timing, deadlines satisfaction, throughput, reliability, energy efficiency and quality of service. The workload, connectivity, and availability of the resources can fluctuate in intensity, thus adaptive scheduling is necessary. Under different edge-cloud scenarios, the intelligent-agent-based scheduling can dynamically coordinate the execution of workloads (Lim, 2024). Facing this challenge, the architectures that work well must continuously monitor and reconfigure their resources to satisfy the application's performance requirements as operation conditions change (Table 1).

Table 1. Comparison of Edge, Fog, and Cloud Computing Paradigms

Paradigm	Key Characteristics	Major Advantages	Limitations	Representative References
Edge	Processing close to	Very low latency;	Limited	Shanmugam et al.

Computing	end devices; decentralized resources	context awareness; reduced data transmission	computation and storage	(2023)
Fog Computing	Intermediate computing between edge and cloud	Regional coordination; distributed processing	Management complexity	Hossain et al. (2023)
Cloud Computing	Centralized high- capacity infrastructure	Scalability; extensive computation and storage	Network dependency; higher latency	Lekkala (2024)
Edge- Cloud	Integrated distributed computing continuum	Combines responsiveness and scalability	Complex orchestration	Almurshed et al. (2025)

3. Resource Allocation

Resource allocation is crucial to adaptive edge-cloud computing, in which heterogeneous computing, networking and storage resources need to be dynamically coordinated under varying workloads and application demands. Allocation of computing resources includes assigning CPUs, memory, GPUs, and AI accelerators based on workload complexity, deadlines, and available resources. The scheduling strategy using deep reinforcement learning can dynamically allocate the workload of AI inference to the edge and cloud resources, which enhances the responsiveness of AI inference than the static scheduling strategy (Ma & Fan, 2025).

Network-resource allocation allocates bandwidth, communication channels, routing and transmission priorities on distributed infrastructure. Communication delays may counterbalance the computational advantages of offloading and resource decisions need to take into account the link condition, traffic variation, and network congestion. In the edge-enabled 5G network scenario, intelligent traffic-adaptive allocation has proven the benefit of dynamically coordinating resources to respond to the varying demands of the network (Chen et al., 2019). The other key aspect is storage allocation, as edge nodes offer low latency access but limited capacity whereas cloud platforms offer scalable storage with larger communication overheads. The data and resource placement in IoT-edge-cloud environment can be improved by utilizing learning-based content distribution (Cui et al., 2023).

Task Offloading is the process of deciding whether a workload should be performed locally, at a nearby edge node, or in the cloud based on the processing requirements, the latency, the connectivity and the energy constraint. The energy efficiency can be enhanced through dynamic coordination of computations and virtualized resources, facilitated by Edge-AI-supported offloading (Hasan et al., 2025). Workload placement decisions are very similar to where applications, services, containers, and dependent tasks should run. Workloads can be redistributed and/or migrated to other services due to changing resource availability and demand. To adapt to these variations, deep-reinforcement-learning-based joint task allocation and service migration is provided (Chen et al., 2022).

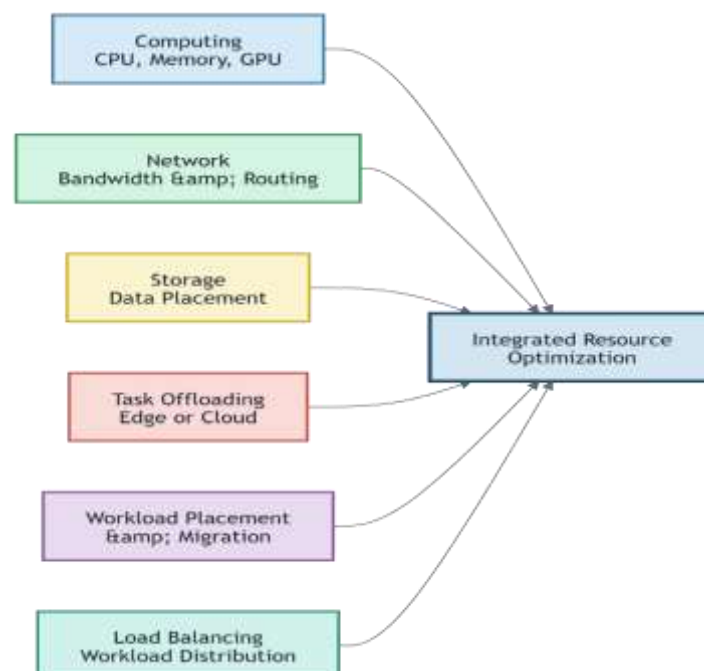


Figure 2. Integrated resource-allocation dimensions in adaptive edge-cloud computing

Load balancing further spreads out the workload between nodes to avoid congestion and under-use. Adaptive learning-based mechanisms can adapt workload distribution based on the real-time system feedback to enhance the utilization and responsiveness (Ghorbani et al., 2025; Ramesh, 2025). In the end, these allocation decisions have to be optimized together; computing capacity, bandwidth, storage space, energy use, latency and cost are tightly coupled (Ajax et al., 2025). To address the conflicting demands and adaptively determine resource configurations in edge-cloud workflows, multi-objective deep reinforcement learning (MODRL) offers a potential solution (Lahza et al., 2024; Anand & Karthikeyan, 2025). Integrated optimization of computing, communication, storage, offloading, placement and load-balancing decisions is required to make effective resource allocation decisions, rather than optimizing each resource separately.

Table 3. Resource-Allocation Strategies in Edge-Cloud Computing

Strategy	Principle	Advantages	Challenges	Representative References
Static	Predefined resource assignment	Simple implementation	Poor adaptability	Shanmugam et al. (2023)
Heuristic	Rule/search-based optimization	Fast decisions	May be suboptimal	Ramesh (2025)
ML-Based	Data-driven prediction and allocation	Predictive management	Training-data dependency	Lilhore et al. (2025)
RL-Based	Reward-driven sequential decisions	Handles dynamic environments	Training instability	Ajax et al. (2025)
DRL-Based	Deep policy/value learning	Handles large state spaces	High computational cost	Anand & Karthikeyan (2025)
Multi-Agent	Distributed intelligent agents	Supports decentralization	Agent coordination	Zhang et al. (2024)
Federated	Distributed collaborative models	Privacy-aware management	Communication overhead	Ala'a et al. (2025)

4. Machine Learning Methods

Machine learning can help make intelligent resource allocation decisions in edge-cloud environments, by learning workload patterns and predicting resource needs, and by enabling dynamic decision making. Various ML paradigms can be employed to solve different aspects of resource management depending on the available data, complexity of the infrastructure and the goals of the allocation. In supervised learning, past data (with known workloads and resource usage) is used to forecast resource usage and workloads, so that resources can be allocated in advance. It proves to be very suitable for resource demand prediction, but it is also influenced by the representative labeled data (Xu & Xu, 2024). Unsupervised learning is able to discover patterns and clusters from unlabeled operational data, and can be used for workload classification, resource usage analysis, though the challenges of validation of discovered patterns still exist (Djigal et al., 2022).

The complex nonlinear relationships between workload properties, network conditions, and resources' state can be captured with deep learning, which is well suited for high-dimensional edge-cloud environments. Neural-network based approaches can boost adaptive resource allocation, but may face the challenge of computational requirements on resource constrained devices at the edge (Almaini et al., 2025). Reinforcement Learning (RL) is well suited to sequential resource-allocation problems, since the agents learn allocation policies from (interacting with) the evolving environment. Another capability of deep reinforcement learning is decision-making in large state spaces for task placement, offloading, and resource assignment (Wang et al., 2019). Multi-agent learning is a technique that distributes decision making duties among several intelligent agents, thereby enabling edge nodes and other resources to coordinate allocation without being totally controlled by a central controller. This has both advantages of scalability and decentralization but also comes with coordination overhead (Zhang et al., 2024). In distributed settings such as the edge and Industrial IoT, federated learning allows for the training of models to be done without sharing local data between the nodes, which helps maintain privacy. However, frequent exchanges of models can lead to increased communication overhead (Ala'a et al., 2025).

Graph neural networks (GNNs) enable topology-aware learning, as they model devices, edge nodes, cloud servers, and their communication patterns as graphs. This capability may be useful in resource allocation, where the decisions are based on relationships between distributed resources, but can be limited by the complexity of the models and thus could not be implemented in real time. (Djigal et al., 2022) In summary, supervised learning and unsupervised learning facilitate prediction and pattern discovery, deep learning models the complex relationships between resources, RL makes adaptive sequential decisions, multi-agent learning achieves decentralized coordination, federated learning allows for privacy-preserving collaboration,

and GNNs learn about the topology of infrastructure. These complementary strategies offer the learning base to intelligent edge–cloud resource allocation as summarized in Table 2.

Table 2. Machine Learning Methods for Resource Allocation

ML Method	Primary Role	Strength	Limitation	Representative References
Supervised Learning	Workload/resource prediction	High predictive capability	Requires labeled data	Xu & Xu (2024)
Unsupervised Learning	Pattern discovery and clustering	Does not require labels	Difficult validation	Djigal et al. (2022)
Deep Learning	Complex workload modeling	Captures nonlinear patterns	Computational overhead	Almaini et al. (2025)
Reinforcement Learning	Sequential allocation	Adaptive decision-making	Convergence and exploration issues	Wang et al. (2019)
Multi-Agent Learning	Distributed allocation	Decentralized decision-making	Coordination overhead	Zhang et al. (2024)
Federated Learning	Collaborative distributed learning	Privacy preservation	Communication overhead	Ala'a et al. (2025)
Graph Neural Networks	Topology-aware learning	Captures resource relationships	Model complexity	Djigal et al. (2022)

5. Adaptive Resource Management

Adaptive resource management combines continuous system awareness, predictive intelligence and dynamic decision making for efficient operation when facing dynamic workloads and infrastructure. Adaptive management adapts resource decisions based on the measured and predicted states of the system, enabling a feedback loop for real-time edge–cloud orchestration that is not found with static allocation. Resource monitoring is the foundation of operations, and it sees data on CPU and memory usage, network bandwidth, queue length, energy usage, application demand, and node availability gathered on an ongoing basis. These observations allow for fast detection of congestion, resource deficiency, and performance degradation and help intelligent controllers help adjust allocation decisions as conditions change.

Workload prediction is an extension of this ability that predicts the demands for computational resources prior to the arrival of resource scarcity. Patterns of tasks and metrics gathered in real-time can be fed into ML models that can predict traffic intensity, application requirements and task arrivals based on historical workload patterns. Predictive resource management also helps make adjustments to workloads proactively, instead of just reactively allocating resources (Lekkala, 2024). Resource forecasting then determines the necessary computing, network, storage, and energy resources to serve expected workloads. In large-scale networks with fluctuating demands and limited provisioning resources, proactively preparing resources based on machine learning prediction and dynamic scheduling can be beneficial (Ying, 2025).

Intelligent task offloading classifies the workload to be run locally, edge nodes or cloud infrastructure depending on the complexity of the task, latency requirements, network conditions, energy requirements, and available processing power. Reinforcement learning is well applicable due to the fact that offloading decisions carry impact on the state of the following resources, which allows the adaptive policies to adapt to mobility and varying computational environments (Zhao et al., 2023). Offloading and dynamic provisioning are complementary concepts and the latter involves the activation, scaling, reallocation, or release of computational resources based on the current or forecasted resource demand. RL-based resource management can adapt resource allocation policies for real-time AI workloads on an ongoing basis, minimizing resource shortages and overprovisioning (Vedula et al., 2025).

With service migration, the app components or execution state are moved when the current deployment locations are inefficient due to congestion, mobility, and resource availability. Efficient migration involves striking a balance between the possible performance gains and migration latency and communication overhead. Adaptive scheduling across heterogeneous edge–cloud resources allows for workload re-distribution with changes of infrastructure (Zhang et al., 2025). Closed-loop adaptation combines these features in a cycling control process where previous allocation outcomes affect future allocation decisions. Hybrid edge–cloud ML pipelines may enable continuous interaction between distributed processing and real-time decision making, enabling the evolution of allocation policies with application conditions and infrastructure, respectively (Asro et al., 2026). The adaptive cycle can be represented as:

Monitor → Predict → Decide → Allocate → Evaluate → Adapt

This feedback-driven cycle provides the foundation for autonomous edge-cloud resource management by continuously observing system conditions, forecasting demand, selecting and executing allocation decisions, evaluating their effects, and adapting subsequent policies to maintain real-time performance.

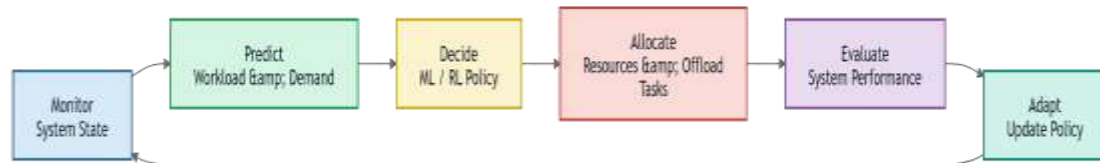


Figure 2. Closed-loop adaptive resource management cycle for intelligent edge-cloud computing

6. Performance Optimization

Optimizing performance of adaptive edge-cloud systems involves optimizing for multiple and often competing objectives, not a single performance measure. Machine-learning-driven allocation allows the resource controller to adapt to changing workload and infrastructure requirements dynamically, while taking into account other factors such as latency, energy use, resource usage, throughput, operational cost, quality of service (QoS), service-level agreements (SLAs) and reliability.

For real-time applications, latency is of special concern since it dictates the responsiveness of the system, as transmission, computation, queuing and migration delays are all included in the overall latency. Intelligent allocation can reduce the end-to-end latency by using specific locations for execution based on workload urgency and availability of resources. For applications that require high responsiveness, like cloud gaming, in which edge resources can have a significant impact, deep reinforcement learning (DRL)-based resource allocation has proved to be appropriate (Deng et al., 2022). It is also vital to have energy efficiency for resource-constrained edge nodes and mobile devices. To achieve energy-aware coordination and ensure the scalability and adaptability of distributed networks, cloud-edge architectures combining digital twins and ML can help (Han et al., 2025).

Efficient resource utilization aims to ensure that the CPU, memory, storage and networking resources are used productively, without being simultaneously overloaded and underutilized. The DRL-based dynamic allocation may allocate resources in real-time based on the system states and the workload requirements (Sulistyo & Setiawan, 2025). Note that throughput is related to utilization: if one scheduling is really good, it has to maximize the number of tasks completed, or the amount of data processed; it must also avoid computational and communication bottlenecks. A reinforcement-learning approach to resource management for real-time AI workloads can dynamically allocate available resources to ensure efficient processing in the face of fluctuating demands (Mallikarjunaradhya et al., 2024).

Cost optimization is also vital since there are differences in the operational costs of cloud execution, communication, resource activation and migration. ML-based allocation can provide a way to determine resource configurations that optimize computational performance and economic efficiency (Bodra & Khairnar, 2025). The application level constraints are defined by QoS and SLA requirements, expressed in terms of latency, throughput, availability and responsiveness. These service requirements can be adapted to changing operational conditions by adjusting the resource allocation decisions, especially for dynamic industrial environments (Vimal et al., 2020). Continuous service even in the event of failures, congestion, connectivity changes, and workload variability also are important aspects of reliability. Distributed hybrid edge-cloud architectures can enhance resilience by offering alternative processing sites, thereby minimizing reliance on specific computational nodes (Ganesh et al., 2025).

These goals are tightly linked together and often there are trade-offs. Edge execution can lower latency, and increase local energy use, while cloud execution can increase computing power, but increase transmission delay and communication cost (Patel, 2025). Likewise, having maximum throughput may result in greater resource contention and thus adversely impact QoS or reliability. Adaptive decision-making (Dong et al., 2021) can be used as a tool to handle such conflicting operational and economic goals through intelligent cloud-edge resource coordination. Multi-objective optimization is thus essential for real-time management of resources between the edge and the cloud, where allocation policies can make the best use of limited resources to meet the latency requirements, energy efficiency, utilization, throughput, cost, QoS, and reliability of each application based upon its priority and the state of the system.

Table 4. Performance Metrics for Resource Allocation

Metric	Description	Measurement	Goal	Representative References
Latency	End-to-end response delay	ms	Minimize	Deng et al. (2022)
Throughput	Tasks/data processed over time	Tasks/s; Mbps	Maximize	Mallikarjunaradhya et al. (2024)

Resource Utilization	Efficiency of available resources	%	Maximize	Sulistyo & Setiawan (2025)
Energy Efficiency	Energy required for processing	J/task; Wh	Optimize	Han et al. (2025)
Cost	Operational/resource expenditure	Monetary cost	Minimize	Patel (2025)
QoS/SLA	Compliance with service requirements	Violation rate (%)	Optimize	Vimal et al. (2020)
Reliability	Successful service operation	Success/availability (%)	Maximize	Ganesh et al. (2025)
Migration Overhead	Cost of workload/service movement	Time; bytes; energy	Minimize	Chen et al. (2022)

7. Applications And Evaluation

Adaptive edge-cloud resource allocation enables applications with low latency, distributed processing and dynamic resource availability. For IoT and Industrial IoT (IIoT), intelligent scheduling is in charge of dynamically assigning workloads to the devices, edge nodes, and cloud-based servers based on the fluctuating computational needs. (Zhang et al., 2025) The transportation, surveillance, utilities, and environmental monitoring aspects of smart cities also rely heavily on edge-cloud integration, where local processing enhances responsiveness, and cloud computing offers scalable analytics (Trigka & Dritsas, 2025). Distributed processing of smart-city IoT data can be further enhanced using collaborative deep learning (Wu et al., 2020).

The network topology and computational requirements are continuously changing in the connected vehicle network, necessitating adaptive resource allocation. The energy-aware edge-cloud allocation scheme can support real-time urban traffic prediction while maintaining resource consumption control (Ali et al., 2025). In healthcare applications, timely processing of sensor data is also crucial, and ML-based adaptive allocation can dynamically adjust to fluctuating data rates and resource demands in healthcare IoT systems (Rafique, 2025). Edge Processing also has its applications in AR/VR and UAV, which, in both cases, demand ultra-fast responses, under the bandwidth, computing, mobility, or energy constraint. Meanwhile, 5G/6G and multi-access edge computing offer communication infrastructures that enable distributed intelligence and low latency orchestration of resources.

Typical evaluations make use of simulation-based environments like CloudSim, iFogSim, or EdgeCloudSim or experiment-based environments using containers or Kubernetes. The typical metrics used for measuring performance are latency, throughput, energy use, resource utilization, cost, task completion, SLA violations, communication overhead, and migration overhead. The accuracy of the predictions, the latency of the inferences, the convergence and the stability of the policy are also to be evaluated by ML-based approaches. DRL-based content distribution highlights the significance of learning driven allocation under application-specific edge-cloud scenario (Neelakantan et al., 2024). Standardized workloads, common baselines and reproducible evaluation procedures are therefore critical to enable meaningful comparisons, and to assess whether reported improvements can be generalized to real-world deployments.

8. Challenges And Future Directions

Despite advances in ML-driven edge-cloud computing, resource heterogeneity remains challenging because devices differ substantially in processing capacity, memory, accelerators, connectivity, and energy availability. There are also dynamic workloads to consider, as well as scalability, and the need to allocate when the demand, mobility and network changes constantly. For future systems, lightweight learning mechanisms that can adapt across heterogeneous and large-scale infrastructures are required (Venkataraman & Ramachandraprabu, 2025). The communication overhead of telemetry, model updates and task migration should also be kept at a minimum to retain the latency benefits of edge.

Security and privacy are crucial as there are potential attacks such as adversarial inputs, compromised nodes, and manipulated model updates in the distributed ML controllers. At the same time, the lack of generality and model drift may impact the rate of allocation if the operating conditions vary from those during training. It is therefore essential that continual & online learning enable real-time adaptation, and explainable AI should aid in making decisions about resource allocation transparent, especially for critical applications.

New technologies are critical for meeting these limitations. Digital twins can be used to simulate the state of resources and test resource allocation policies prior to deployment; Green edge computing can integrate energy cost, carbon intensity, and availability of renewable energy into resource scheduling. There are already several examples of how the digital-twin and ML integration is a potential for scalable and energy-efficient cloud-edge coordination (Han et al., 2025). The integration of communication, computation, sensing, and intelligence may be further achieved by AI-native 6G, which enables highly responsive distributed services. The long-term direction is towards autonomous orchestration, where edge-cloud systems continuously observe conditions, forecast demand, allocate resources, assess performance and autonomously make optimisations to

their policies. Support for scalable continual learning, secure coordination, explainable decisions, energy efficient optimization, and strong generalization for reliable self-managing edge–cloud infrastructures should therefore be a focus of future research efforts.

9. Conclusion

Edge–cloud integration with machine learning is a big step towards intelligent and responsive resource management. This review shows that a coordinated infrastructure of both edges and clouds is very useful to overcome the drawbacks of having a purely centralized or localized computing infrastructure that statically allocates computation, communication and storage resources to meet some application needs. This capability is supported by ML-driven methods, including workload prediction, intelligent task offloading, dynamic provisioning, service migration and adaptive scheduling, among others, notably deep learning, reinforcement learning, multi-agent learning and federated learning. These approaches must be effectively coordinated across heterogeneous resources and must have the capability of balancing latency, energy usage, resource utilization, throughput, cost, QoS and reliability. But, in spite of all these advances, the deployment of real-world applications is still constrained by the challenges of resource heterogeneity, dynamic workloads, communication overhead, scalability, security and privacy risks, model drift, limited generalization, and learning overhead. This will need lightweight and explainable learning, powerful distributed intelligence, continuous adaptation, and uniformity in assessment tools. Further, new trends in digital twins, green computing, and AI-enabled 6G offer insights into even more predictive and sustainable management of resources. In the end, adaptive edge–cloud architectures look to become autonomous orchestration systems, capable of continuously monitoring, predicting, allocating, evaluating and optimizing resources with minimal human intervention, and that will serve as the foundation for scalable and secure, energy-efficient and real-time intelligent computing.

References

1. Ajax, R., Wasiiu, R., & Daniel, S. (2025). Deep Reinforcement Learning for Resource Allocation in Edge-Cloud Environments. Preprint.
2. Ala'a, R., Karim, F. K., & Wang, Y. (2025). Optimizing resource allocation in industrial IoT with federated machine learning and edge computing integration. *Results in Engineering*, 106387.
3. Ali, A., Ullah, I., Singh, S. K., Sharafian, A., Jiang, W., I. Sherazi, H., & Bai, X. (2025). Energy-efficient resource allocation for urban traffic flow prediction in edge-cloud computing. *International Journal of Intelligent Systems*, 2025(1), 1863025.
4. Almaini, A., Folz, J., Al-Dubai, A., Schramm, M., Heigl, M., & Ghaleb, B. (2025, July). Adaptive resource allocation in edge computing for iot networks using neural networks. In *2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)* (pp. 514-521). IEEE.
5. Almurshed, O., Kaushal, A., Meshoul, S., Muftah, A., Almoghamis, O., Petri, I., ... & Rana, O. (2025). Enhancing performance of machine learning tasks on edge-cloud infrastructures: A cross-domain Internet of Things based framework. *Future Generation Computer Systems*, 166, 107696.
6. Anand, J., & Karthikeyan, B. (2025). EADRL: Efficiency-aware adaptive deep reinforcement learning for dynamic task scheduling in edge-cloud environments. *Results in Engineering*, 27, 105890.
7. Asro, A., Solihin, S., & Irlon, I. (2026). Optimizing End to end Machine Learning Pipelines Using Hybrid Edge Cloud Architectures for Real Time Decision making Applications. *Big Data Analytics and Data Science*, 1(1), 35-44.
8. Bataineh, J. K., & Tawalbeh, R. M. (2026, May). Adaptive Resource Allocation in Edge Computing Using Intelligent System Response. In *2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)* (pp. 1-5). IEEE.
9. Bodra, D., & Khairnar, S. (2025). Machine learning-based cloud resource allocation algorithms: a comprehensive comparative review. *Frontiers in computer science*, 7, 1678976.
10. Chen, M., Miao, Y., Gharavi, H., Hu, L., & Humar, I. (2019). Intelligent traffic adaptive resource allocation for edge computing-based 5G networks. *IEEE transactions on cognitive communications and networking*, 6(2), 499-508.
11. Chen, Y., Sun, Y., Wang, C., & Taleb, T. (2022). Dynamic task allocation and service migration in edge-cloud IoT system based on deep reinforcement learning. *IEEE Internet of Things Journal*, 9(18), 16742-16757.
12. Cui, T., Yang, R., Fang, C., & Yu, S. (2023). Deep reinforcement learning-based resource allocation for content distribution in IoT-edge-cloud computing environments. *Symmetry*, 15(1), 217.
13. Deng, X., Zhang, J., Zhang, H., & Jiang, P. (2022). Deep-reinforcement-learning-based resource allocation for cloud gaming via edge computing. *IEEE Internet of Things Journal*, 10(6), 5364-5377.
14. Djigal, H., Xu, J., Liu, L., & Zhang, Y. (2022). Machine and deep learning for resource allocation in multi-access edge computing: A survey. *IEEE Communications Surveys & Tutorials*, (4), 2449-2494.

15. Dong, W., Yang, Q., Li, W., & Zomaya, A. Y. (2021). Machine-learning-based real-time economic dispatch in islanding microgrids in a cloud-edge computing environment. *IEEE Internet of Things Journal*, 8(17), 13703-13711.
16. Ganesh, G., Nasreen, M. A., & Jaganathan, A. (2025, December). Hybrid Edge-Cloud AI and Blockchain System Architecture for Real-Time Decision-Making in IT Project Management. In *2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 751-755). IEEE.
17. Ghorbani, M., Khaledian, N., & Moazzami, S. (2025). ALBLA: an adaptive load balancing approach in edge-cloud networks utilizing learning automata: M. Ghorbani et al. *Computing*, 107(1), 34.
18. Han, L., Lei, M., He, G., Li, Y., & Zhao, Y. (2025). Energy-efficient cloud-edge collaborative model integrating digital twins and machine learning for scalable and adaptive distributed networks. *Sustainable Computing: Informatics and Systems*, 47, 101157.
19. Hasan, R. A., Yasin, K., Kareem, P. R., Salih, A. M., Jasim, H., Ameerudeen, M. A., ... & Rachini, A. (2025). Energy-efficient task offloading and resource allocation in mobile cloud computing using edge-AI and network virtualization. *Khwarizmia*, 2025, 42-49.
20. Hossain, M. E., Tarafder, M. T. R., Ahmed, N., Al Noman, A., Sarkar, M. I., & Hossain, Z. (2023). Integrating AI with edge computing and cloud services for real-time data processing and decision making. *International journal of multidisciplinary sciences and arts*, 2(4), 252-261.
21. Khan, F. U., Shah, I. A., Jan, S., Ahmad, S., & Whangbo, T. (2025). Machine Learning-Based Resource Management in Fog Computing: A Systematic Literature Review. *Sensors*, 25(3), 687. <https://doi.org/10.3390/s25030687>
22. Lahza, H., BR, S., & Lahza, H. F. M. (2024). Adaptive multi-objective resource allocation for edge-cloud workflow optimization using deep reinforcement learning. *Modelling*, 5(3), 1298-1313.
23. Lekkala, C. (2024). Ai-driven dynamic resource allocation in cloud computing: Predictive models and real-time optimization. *J Artif Intell Mach Learn & Data Sci*, 2.
24. Lilhore, U. K., Simaiya, S., Sharma, Y. K., Rai, A. K., Padmaja, S. M., Nabilal, K. V., ... & Alsufyani, H. (2025). Cloud-edge hybrid deep learning framework for scalable IoT resource optimization. *Journal of Cloud Computing*, 14(1), 5.
25. Lim, J. (2024). Adaptive scheduling based on intelligent agents in edge-cloud computing environments. *Journal of Internet Technology*, 25(4), 609-617.
26. Ma, X., & Fan, S. (2025). Adaptive Scheduling Algorithm for AI Inference Tasks Based on Deep Reinforcement Learning in Cloud-Edge Collaborative Environment. *Annals of Applied Sciences*, 6(1).
27. Mallikarjunaradhya, V., Sreeramulu, M. D., Mohammed, A. S., Boddapati, N., Gupta, K., & Natarajan, Y. (2024, September). Efficient resource management for real-time AI systems in the cloud using reinforcement learning. In *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)* (Vol. 7, pp. 1654-1659). IEEE.
28. Neelakantan, P., Gangappa, M., Rajasekar, M., Sunil Kumar, T., & Suresh Reddy, G. (2024). Resource allocation for content distribution in IoT edge cloud computing environments using deep reinforcement learning. *Journal of High Speed Networks*, 30(3), 409-426.
29. Patel, H. K. (2025). Adaptive resource allocation using reinforcement learning for performance and cost optimization. *Journal of Computer Science and Technology Studies*, 7(9), 688-694.
30. Rafique, W. (2025). ML-raspf: A machine learning-based rate-adaptive framework for dynamic resource allocation in smart healthcare iot. *Algorithms*, 18(6), 325.
31. Ramesh, R. (2025). EdgeBrain: Real-Time AI with Modified Firefly Algorithms in Edge Computing. *Artificial Intelligence and Society*, 104.
32. Shah, S. C. (2021). Design of a machine learning-based intelligent middleware platform for a heterogeneous private edge cloud system. *Sensors*, 21(22), 7701.
33. Shanmugam, L., Jangoan, S., & Sharma, K. K. (2023). Dynamic resource allocation in edge computing for ai/ml applications: Architectural framework and optimization techniques. *Journal of Knowledge Learning and Science Technology* ISSN: 2959-6386 (online), 2(2), 385-397.
34. Shen, W., Lin, W., Wu, W., Wu, H., & Li, K. (2025). Reinforcement learning-based task scheduling for heterogeneous computing in end-edge-cloud environment. *Cluster Computing*, 28(3), 179.
35. Sıcakyüz, Ç., Asaad, R. R., Almufti, S., & Rustamova, N. R. (2024). Adaptive Deep Learning Architectures for Real-Time Data Streams in Edge Computing Environments. *Qubahan Techno Journal*, 3(2), 1-14.
36. Sulistyo, M. A., & Setiawan, D. (2025). Deep reinforcement learning-based algorithm for dynamic resource allocation in edge computing. *ALCOM: Journal of Algorithm and Computing*, 1(1), 13-22.
37. Trigka, M., & Dritsas, E. (2025). Edge and cloud computing in smart cities. *Future Internet*, 17(3), 118.
38. Tuli, S., Ilager, S., Ramamohanarao, K., & Buyya, R. (2020). Dynamic scheduling for stochastic edge-

- cloud computing environments using a3c learning and residual recurrent neural networks. *IEEE transactions on mobile computing*, 21(3), 940-954.
39. Vedula, J., Baskaran, S., Sayiram, S., Muthusamy, R., Gupta, K., & Yuvaraj, N. (2025, February). Efficient resource management for real-time AI systems in the cloud using reinforcement learning. In 2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-6). IEEE.
 40. Venkataraman, H., & Ramachandraprabu, S. P. (2025). A Hybrid Deep Learning and Edge Computing Framework for Real-Time IoT Data Processing. *Journal of Computer Applications and Information Technology*.
 41. Vimal, S., Khari, M., Dey, N., Crespo, R. G., & Robinson, Y. H. (2020). Enhanced resource allocation in mobile edge computing using reinforcement learning based MOACO algorithm for IIOT. *Computer Communications*, 151, 355-364.
 42. Wang, J., Zhao, L., Liu, J., & Kato, N. (2019). Smart resource allocation for mobile edge computing: A deep reinforcement learning approach. *IEEE Transactions on emerging topics in computing*, 9(3), 1529-1541.
 43. Wu, H., Zhang, Z., Guan, C., Wolter, K., & Xu, M. (2020). Collaborate edge and cloud computing with distributed deep learning for smart city internet of things. *IEEE Internet of Things Journal*, 7(9), 8099-8110.
 44. Xu, G., & Xu, M. (2024). An effective prediction of resource using machine learning in edge environments for the smart healthcare industry. *Journal of Grid Computing*, 22(2), 54.
 45. Ying, G. (2025, December). Research on a Machine Learning and Cloud Computing-Based System for Real-Time Prediction, Fast Decision-Making, and Dynamic Resource Scheduling in Large-Scale Networks. In 2025 IEEE 4th International Conference of Safe Production and Informatization (IICSPI) (pp. 558-564). IEEE.
 46. Zhang, T., Wu, F., Chen, Z., & Chen, S. (2024). Optimization of edge-cloud collaborative computing resource management for Internet of Vehicles based on multiagent deep reinforcement learning. *IEEE Internet of Things Journal*, 11(22), 36114-36126.
 47. Zhang, Y., Zhao, X., Liu, M., Liu, J., & Zheng, R. (2025). A cloud edge resources scheduling method based on reinforcement learning in industrial internet of things. *Cluster Computing*, 28(15), 963.
 48. Zhao, J., Quan, H., Xia, M., & Wang, D. (2023). Adaptive resource allocation for mobile edge computing in internet of vehicles: A deep reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 73(4), 5834-5848.