frontiers in the Internet of Things

TYPE Perspective
PUBLISHED 27 August 2026
DOI 10.3389/friot.2026.1883673

OPEN ACCESS

EDITED BY
Minxian Xu,
Chinese Academy of Sciences (CAS),
China

REVIEWED BY
Ibrahim Algahmdi,
Al Baha University, Saudi Arabia

*CORRESPONDENCE
Zhiyu Wang,
zhiywang1@student.unimelb.edu.au

RECEIVED 17 May 2026
REVISED 17 June 2026
ACCEPTED 17 June 2026
PUBLISHED 27 August 2026

CITATION
Wang Z, Kapri N, Bittencourt LF and Buyya R (2026) Artificial intelligence-driven resource management in edge-cloud computing continuum for internet of things applications: new trends and future directions.
Front. Internet Things 5:1883673.
doi: 10.3389/friot.2026.1883673

Artificial intelligence-driven resource management in edge-cloud computing continuum for internet of things applications: new trends and future directions

Zhiyu Wang[1]*, Nilotpal Kapri[1,2,3], Luiz Fernando Bittencourt[4] and Rajkumar Buyya[1]

[1]Quantum Cloud Computing and Distributed Systems (qCLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia, [2]Electronics, Instrumentation and Control Department, CSIR–Institute of Minerals and Materials Technology, Bhubaneswar, India, [3]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India, [4]Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil

The rapid proliferation of Internet of Things (IoT) devices across domains such as smart transportation, healthcare, and smart city infrastructure has intensified the demand for low-latency, energy-efficient, and scalable computing paradigms. While cloud computing has traditionally served as the backbone for IoT data processing, its inherent limitations have catalyzed the emergence of edge/fog computing, forming a distributed edge-cloud continuum. This review examines emerging trends in Artificial Intelligence (AI)-driven resource management within this continuum, with a focus on three directions: (1) the transition from centralized to distributed and collaborative intelligence, (2) cross-domain adaptation and knowledge transfer for heterogeneous IoT applications, and (3) the nascent integration of foundation models into edge environments. We ground our discussion in three complementary case studies: smart transportation as a representative vertical domain, cross-domain heterogeneous IoT application scheduling as a horizontal perspective, and industrial transferred arc plasma monitoring as an emerging industrial IoT scenario, and conclude with forward-looking research directions, including quantum-enhanced edge optimization, edge-native continual learning, digital twin-driven resource orchestration, and neuromorphic computing for ultra-low-power edge AI.

KEYWORDS
artificial intelligence, cloud computing, edge computing, internet of things, resource management

1 Introduction

The Internet of Things (IoT) has fundamentally reshaped the way physical and digital worlds interact. The number of globally active IoT devices reached 17.7 billion at the end of 2024 and is projected to grow to 40.6 billion by 2034 (Transforma Insights, 2025), generating unprecedented volumes of data that must be processed in real time. Large-scale IoT deployments now underpin critical applications in smart transportation (Kulandaivel and Akuthota, 2026), remote healthcare monitoring, smart city applications, and intelligent building management.

These applications share stringent computational requirements: ultra-low latency, high reliability, and the ability to process massive data streams near the source. Traditionally, cloud computing has served as the primary computational backbone, offering scalable storage and processing in centralized data centers. However, transmitting IoT data to remote servers introduces significant latency, consumes substantial bandwidth, concentrates data storage, and raises privacy concerns, particularly in data-sensitive, delay-constrained scenarios such as autonomous driving (Cao et al., 2024).

These limitations have motivated the emergence of edge/fog computing as complementary paradigms to the cloud, bringing computation closer to the data source (Cao et al., 2024). Rather than replacing cloud infrastructure, these paradigms extend it, forming a distributed edge-cloud continuum that enables flexible, context-aware resource allocation (Srirama, 2024). Within this continuum, a central research challenge is resource management: the intelligent orchestration of computation, storage, communication, and energy resources across heterogeneous and geographically distributed nodes (Vali et al., 2026; Rasouli et al., 2026).

This review examines emerging trends that are reshaping resource management in edge-cloud IoT systems. We trace the architectural evolution of the computing continuum, survey recent advances in Artificial Intelligence (AI)-driven resource management with emphasis on distributed intelligence and cross-domain adaptation, and discuss practical implications through three complementary case studies: smart transportation as a single-domain scenario, cross-domain heterogeneous IoT application scheduling as a multi-domain scenario, and industrial transferred arc plasma monitoring as an emerging industrial IoT scenario. The article concludes with a discussion of open challenges and future research directions.

2 The edge-cloud computing continuum

Cloud computing provides on-demand access to powerful computational resources in centralized data centers. While effective for batch-processing workloads, this model encounters fundamental bottlenecks when applied to modern IoT: network latency ranging from tens to hundreds of milliseconds, unsustainable bandwidth consumption from billions of continuously streaming devices, and regulatory concerns over centralized processing of sensitive data (Cao et al., 2024).

The idea of pushing computation closer to the end user was pioneered by early visions of fog computing (Bonomi et al., 2012) and further advanced through the cross-tier orchestration of resources across the cloud, cloudlets, and devices for edge-native IoT applications (Satyanarayanan et al., 2022). Building on these foundations, edge computing addresses these limitations by positioning computational resources at or near the network periphery, such as on gateways, base stations, or roadside units, thereby reducing latency and enabling local, low-latency data processing (Rasouli et al., 2026; Vetriveeran and Leena Sri, 2025). Rather than viewing edge and cloud as competing alternatives, the contemporary paradigm treats them as an integrated edge–cloud continuum, where workloads are dynamically distributed across tiers based on latency requirements, resource availability, and application context (Srirama, 2024; Goudarzi et al., 2022). Figure 1 illustrates this three-tier architecture, where IoT devices generate data at the periphery, edge/fog nodes provide proximate computation and storage, and cloud instances offer scalable centralized resources. This hierarchical design has become a common architectural paradigm for modern IoT systems (Goudarzi et al., 2022).

The distributed and heterogeneous nature of this continuum introduces complex resource management challenges (Bittencourt et al., 2025). Key tasks include task offloading (deciding where a computation should execute), resource allocation (assigning CPU, memory, and bandwidth to competing tasks), workload scheduling (ordering task execution to meet Quality of Service (QoS) requirements), and service placement (determining where to deploy application components) (Vali et al., 2026; Rasouli et al., 2026). Traditional optimization approaches such as convex programming and heuristics struggle with the scale, dynamism, and uncertainty inherent in these environments. Convex methods require static, well-structured problem formulations that rarely hold under fluctuating edge conditions, while heuristics rely on hand-crafted rules that lack adaptability across diverse scenarios. These limitations have motivated the adoption of AI-driven approaches that can learn adaptive decision policies directly from system interactions (Zhou et al., 2024; Vali et al., 2026).

3 Emerging trends in AI-Driven resource management

3.1 From centralized to distributed and collaborative intelligence

Deep Reinforcement Learning (DRL) has been widely adopted for resource management in edge-cloud environments, as DRL agents can learn decision policies through interaction with dynamic, stochastic environments (Zhou et al., 2024). Early DRL approaches often relied on centralized architectures, where a single agent observes the global system state and makes scheduling or offloading decisions. While effective in small-scale settings, centralized DRL faces scalability bottlenecks, single points of failure, and communication overhead in geographically distributed edge environments (Wang et al., 2026).

To avoid the centralization drawbacks, a major emerging trend is the shift toward distributed and collaborative intelligence. Li et al. (2023) proposed TapFinger, a GNN-based multi-agent reinforcement learning scheduler for edge clusters that co-optimizes task placement and fine-grained resource allocation, achieving up to 54.9% reduction in task completion time compared to state-of-the-art baselines. Chen et al. (2024) formulated the distributed resource management problem as a partially observable Markov decision process (POMDP) and solved it via a two-step multi-agent DRL approach, enabling each edge server to make autonomous task allocation decisions based on

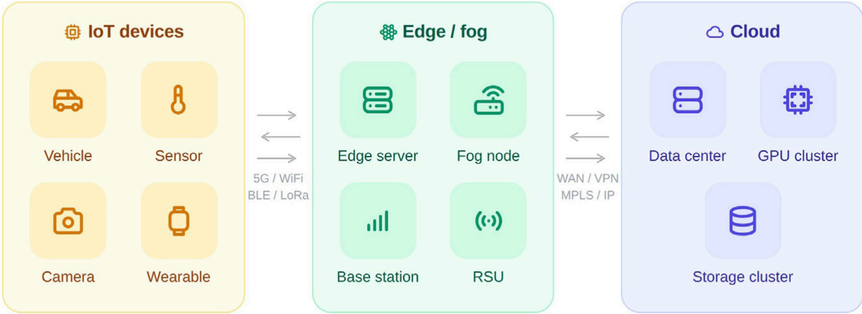

FIGURE 1
The edge-cloud computing continuum.

local observations. Wang et al. (2025c) introduced a transformer-enhanced distributed DRL architecture in which multiple edge nodes collaboratively learn scheduling policies, leveraging attention mechanisms to capture inter-task dependencies at scale. These distributed approaches collectively represent a paradigm shift from monolithic, centralized decision-making toward collaborative, multi-agent intelligence at the network edge. These multi-agents at the edge can cooperate among each other and also with cloud agents that have a higher-level view of the system objectives, but without the need to monitor the whole system with very small granularity as in a fully centralized approach.

3.2 Cross-domain adaptation and knowledge transfer

A second emerging trend addresses a practical limitation of existing DRL methods: domain specificity. A scheduling policy trained for smart transportation workloads may perform poorly when applied to healthcare or industrial IoT scenarios due to differences in workload characteristics, QoS requirements, resource constraints, and performance objectives. Retraining from scratch for each new domain is prohibitively expensive.

Recent work has explored several strategies to enable cross-domain generalization. Meta-Reinforcement Learning (Meta-RL) allows agents to rapidly adapt to new environments with minimal fine-tuning. Tong et al. (2026) applied Meta-RL to cloud-edge-end resource allocation in Intelligent Transportation Systems (ITS), demonstrating rapid adaptation to non-Independent-and-Identically-Distributed (non-IID) traffic pattern shifts. Knowledge distillation offers another pathway, enabling compact student models to inherit decision capabilities from larger teacher models. Wang et al. (2026) proposed a knowledge distillation-empowered adaptive Federated Reinforcement Learning (FRL) framework that enables cross-domain policy transfer while preserving data privacy, demonstrating robust performance across multi-domain IoT applications. These advances signal a shift from single-scenario optimization toward generalizable, transferable edge intelligence. Notwithstanding, managing multiple cross-domain distributed learning setups at the same time and adaptively, reflecting heterogeneous requirements, models and dynamics of each domain, is still a relevant research challenge.

3.3 Foundation models for edge intelligence

A more nascent trend is the exploration of foundation models in edge computing environments. Large Language Models (LLMs) such as GPT, LLaMA, and Qwen have demonstrated strong capabilities in reasoning and decision-making (Zheng et al., 2025). Their potential IoT applications include intelligent data analysis, semantic communication that transmits meaning rather than raw data, and natural language-driven system management (Wang R. et al., 2025; Liang et al., 2026). For resource management specifically, LLMs could serve as high-level orchestrators that interpret application requirements expressed in natural language and translate them into scheduling or allocation policies, and reason over system logs to diagnose resource bottlenecks and recommend reconfigurations (Wang R. et al., 2025). LLMs-driven resource management and auto-scaling decisions for microservice applications deployed in cloud computing environments are effective and, at the same reduced the complexity of algorithmic design and implementation (Bai et al., 2026).

Deploying such models on resource-constrained edge devices remains challenging due to memory, compute, and energy limitations (Wang R. et al., 2025). Techniques such as quantization, knowledge distillation, and pruning can reduce model footprints and computational costs (Zhu et al., 2024; Liang et al., 2026). While still in its early stages, the convergence of foundation models with edge computing represents a potential paradigm shift in how IoT systems process and reason about data. In this scenario, Small Language Models (SLMs) and TinyML models can be utilized for specialized decision-making in cooperation with LLMs running in resource-capable infrastructures for a better understanding of the optimization at system-wide levels. Defining the data flow from applications and devices to SLMs and LLMs, their interaction interfaces and frequency, and the granularity of the decision making at each level are questions that will demand research in the near future.

4 Case studies

4.1 Case study 1: Smart transportation

Smart transportation provides a representative case study for the convergence of edge-cloud computing and AI-driven resource

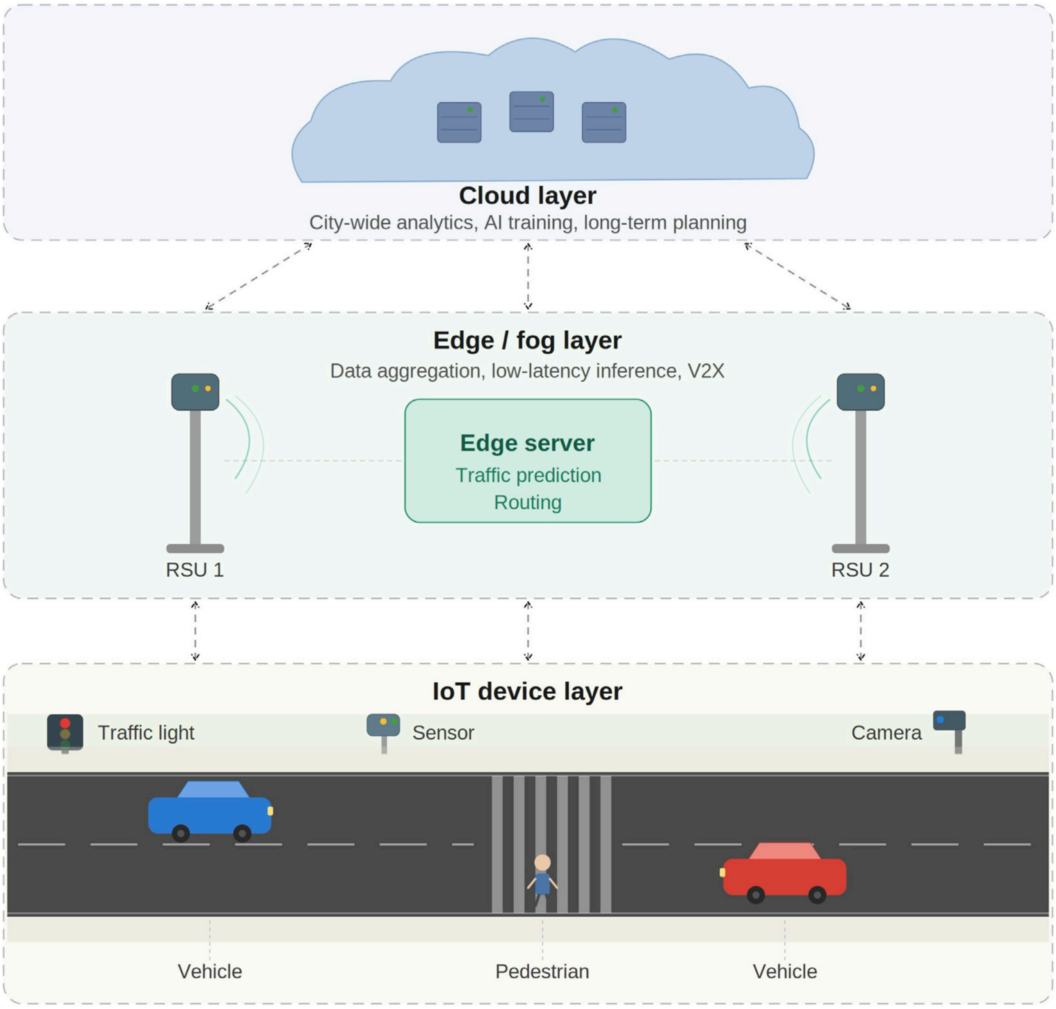

FIGURE 2
AI-driven smart transportation.

management. Modern ITS integrate diverse IoT devices, including connected vehicles, Roadside Units (RSUs), traffic cameras, and environmental sensors, generating massive volumes of heterogeneous, time-sensitive data (Kulandaivel and Akuthota, 2026). Figure 2 depicts a representative smart transportation scenario, where intelligent vehicles and pedestrians interact with RSUs at the edge layer, which in turn coordinates with cloud resources for large-scale analytics and planning.

The edge–cloud continuum serves distinct roles in this context. At the device layer, onboard units perform real-time tasks such as object detection and collision avoidance. At the edge layer, RSUs and regional centers aggregate data for route optimization, traffic flow prediction and optimization (e.g. controlling unsignalized intersections), and city-wide real-time congestion avoidance. At the cloud layer, agencies conduct city-wide analytics and long-term planning.

DRL-based resource management has proven effective in this domain. Joint optimization of computation offloading and power allocation in Vehicular Edge Computing (VEC) has yielded significant latency and energy improvements (Qiu et al., 2024). Multi-agent DRL enables cooperative offloading among RSUs under dynamic vehicular conditions (Fan et al., 2024; Ma et al., 2025), and Meta-RL frameworks allow ITS resource managers to adapt to shifting traffic patterns without retraining (Tong et al., 2026).

4.2 Case study 2: Cross-domain heterogeneous IoT application scheduling

Unlike the smart transportation scenario discussed in Section 4.1, which focuses on a single vertical domain, this subsection takes a horizontal perspective to examine resource management challenges arising from diverse IoT applications spanning multiple domains within the edge-cloud continuum. In practice, edge-cloud systems are often required to simultaneously host heterogeneous applications from different fields, including activity recognition and pose estimation for healthcare monitoring, face detection and video-based optical character recognition for intelligent surveillance, sentiment analysis and speech recognition for natural language processing, and image processing and data compression for general-purpose data services (Wang et al., 2026). These applications exhibit significantly different computational characteristics and dynamics, ranging from I/O-intensive and CPU-intensive workloads to those requiring model inference acceleration or constrained by network bandwidth. They are typically composed of multiple interdependent subtasks forming Directed Acyclic Graphs (DAGs), each with distinct requirements for completion time, energy consumption, and quality of service. Deploying such heterogeneous applications across multi-domain infrastructure comprising cloud servers, edge nodes, and IoT devices introduces three intertwined resource management challenges: substantial disparities in device computational capabilities, non-IID workload patterns across domains, and the need for cross-domain collaborative learning while preserving operational data privacy.

Recent work has begun to address these challenges, directly reflecting the emerging trends identified in Section 3. In the direction of distributed and collaborative intelligence, Wang et al. (2025c) proposed TF-DDRL, which employs an IMPALA-based distributed DRL architecture augmented with Transformer attention mechanisms to capture inter-task dependencies in DAG-structured applications, enabling multiple edge nodes to collaboratively learn scheduling policies and overcoming the scalability and reliability limitations of centralized approaches. In the direction of cross-domain adaptation and knowledge transfer, Wang et al. (2026) proposed KD-AFRL, a framework that integrates three complementary mechanisms: resource-aware adaptive architecture generation through dual-zone neural networks (shared foundational zones and personalized adaptation zones) that enable devices with varying computational capabilities to participate in federated learning; privacy-preserving environment-clustered federated learning that leverages differential privacy and K-means clustering to group environmentally similar domains for targeted collaboration, mitigating performance degradation caused by non-IID distributions; and environment-oriented cross-architecture knowledge distillation that transfers scheduling knowledge between heterogeneous models through temperature-regulated soft targets, allowing resource-constrained devices to benefit from the learning outcomes of more capable domains. Both approaches were validated on the ReinFog platform (Wang et al., 2025b), a modular DRL-based resource management platform for edge-cloud environments that supports both centralized and distributed scheduling techniques and provides a complete experimental pipeline from simulation to deployment on real cloud-edge-IoT infrastructure.

Although these techniques were validated in the context of multi-domain IoT application scheduling, their core mechanisms offer broad applicability. Resource-aware adaptive architecture generation can be extended to any edge deployment scenario involving device heterogeneity. Environment-clustered federated learning provides a paradigm for privacy-preserving collaboration across institutions and geographic regions, such as joint optimization of medical IoT systems across multiple hospitals or coordinated scheduling of manufacturing systems across industrial parks. Cross-architecture knowledge distillation offers a viable pathway for deploying high-quality decision models on resource-constrained edge devices, forming a potential complement to the deployment of foundation models at the edge as discussed in Section 3.3. This case study complements the single-domain perspective of Section 4.1 by demonstrating from the standpoint of application diversity that AI-driven resource management techniques have the potential to generalize beyond specific vertical scenarios toward deployment in cross-domain heterogeneous environments.

4.3 Case study 3: Industrial process (transferred arc plasma) monitoring and control

In aluminium production industries, bauxite ore from mining is chemically processed to extract alumina (aluminium oxide), and red mud is produced as a waste by-product. Transferred Arc Plasma (TAP) systems are widely used in such material processing, metallurgy, and waste management applications to extract valuable metals from red mud-like residues (Jovičević-Klug et al., 2024). However, such existing systems still heavily rely on manual operation through expert supervision. Intelligent monitoring and control capabilities remain largely at the laboratory and pilot-plant stage. Recent advances in soft sensing and fault diagnosis indicated the requirement of future edge-cloud Industrial Internet of Things (IIoT) deployments. Sethi et al. (2023b) demonstrated the use of advanced signal processing techniques to detect instabilities and faults in transferred arc plasma systems through the analysis of multimodal sensor data, including electrical, acoustic, optical, and spectroscopic signals. In a related study, Sethi et al. (2023a) proposed a CNN-based deep learning model that estimates process parameters from plasma images, providing soft-sensing capabilities that can maintain monitoring performance during sensor failures.

Building on these works, a future deployment scenario can be envisioned in which multiple plasma systems in different industrial platforms are connected through an edge-cloud continuum as illustrated in Figure 3. These systems are used for the reduction of industrial wastes such as red mud and to extract valuable resources such as steel and critical minerals. In these systems, there are two electrodes: one is a crucible-type anode that holds the material under process, and another is a cylindrical cathode that controls the power delivery and plasma generation by adjusting the gap between these two electrodes. A channel through the cathode facilitates gas flow to the furnace. Cathode movement is controlled using a Variable Frequency Drive (VFD) connected to a 3-phase asynchronous motor. Multiple sensors such as camera, photodiode, Sound Pressure Level meter (SPL), rotameter, and Linear Variable Differential Transformer (LVDT) can be used for acquiring various process variables (light, sound, gas flow rate, and cathode position) and AI models can be trained on these sensor data. Edge nodes could perform real-time fault detection and process monitoring using these AI models, and adjust cathode position, current level, and gas flow to the furnace for efficient mineral extraction operation, while cloud platforms support model training, large-scale analytics, and knowledge sharing between different systems. From a resource management perspective, latency-sensitive monitoring tasks would be executed at the edge, whereas computationally intensive workloads could be offloaded to the

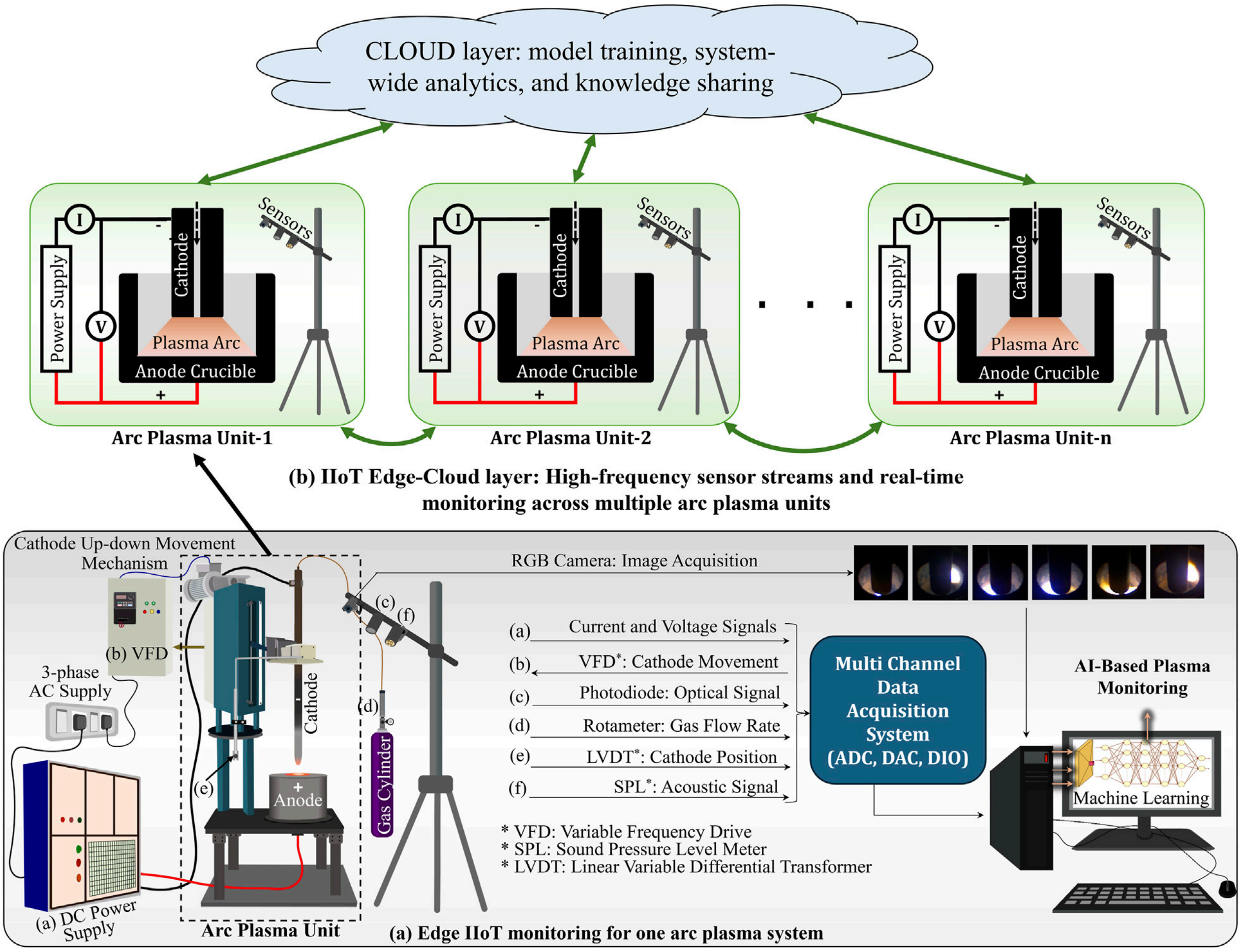

FIGURE 3
Edge-cloud continuum for distributed transferred arc plasma monitoring and control in industrial IoT environments.

cloud. AI-driven resource management techniques could further optimize the allocation of computing, storage, and communication resources across the continuum.

Although fully autonomous arc plasma processing systems have not yet been realized, these studies provide foundational works for future distributed edge-cloud control architectures. This case highlights the potential of AI-enabled monitoring, adaptive resource allocation, and collaborative learning in next-generation IIoT environments.

5 Future directions

Despite significant progress, several open challenges warrant further investigation.

5.1 Quantum-enhanced edge optimization

Combinatorial optimization problems in resource scheduling and task allocation are computationally intractable at scale. Quantum annealing and variational quantum algorithms offer potential speedups for these NP-hard problems. As quantum hardware matures, hybrid quantum-classical frameworks could enable edge orchestrators to solve scheduling problems that are infeasible for classical methods alone. Quantum computing infrastructures can act as accelerators for specific tasks where they are most efficient.

5.2 Edge-native continual learning

Current approaches predominantly follow a train-in-cloud, deploy-at-edge paradigm. However, edge environments are inherently non-stationary: traffic patterns shift with seasons, events, and infrastructure changes. Future systems should support continual learning at the edge to cope with constant concept drifts, enabling models to adapt incrementally without catastrophic forgetting and without requiring costly retraining cycles in the cloud.

5.3 Digital twin-driven resource orchestration

Digital twins can provide high-fidelity virtual replicas of edge-cloud systems, enabling what-if analysis and predictive resource allocation. In smart transportation, a digital twin of the urban road network could simulate congestion scenarios and proactively redistribute computational resources before bottlenecks materialize (Bittencourt et al., 2024).

5.4 Neuromorphic computing for ultra-low-power edge AI

Spiking Neural Networks (SNNs) running on neuromorphic processors consume orders of magnitude less energy than conventional deep learning accelerators. For battery-powered IoT edge devices, neuromorphic architectures could enable always-on AI inference at power budgets that are infeasible with current hardware.

6 Conclusion

The evolution from cloud-centric to edge-cloud computing has transformed the IoT landscape, enabling latency-sensitive, privacy-aware applications across smart transportation, healthcare, and the built environment. This review has examined three emerging trends in AI-driven resource management: the shift from centralized to distributed and collaborative DRL, cross-domain adaptation through knowledge distillation and federated learning, and the nascent integration of foundation models into edge environments. Through three complementary case studies, we have illustrated how these trends manifest in single-domain settings such as smart transportation, multi-domain scenarios involving heterogeneous IoT applications with diverse computational characteristics, and emerging industrial IoT deployments such as transferred arc plasma monitoring and control. Significant challenges remain, spanning quantum-enhanced optimization, edge-native continual learning, digital twin-driven resource orchestration, and neuromorphic computing. Addressing these will require interdisciplinary collaboration across computer science, engineering, and domain-specific fields. As the IoT ecosystem continues to expand, intelligent AI-driven resource management will be essential to realizing the full potential of the edge-cloud computing continuum.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: No datasets were generated or analyzed in this study. This article is a perspective/review and does not involve original data.

Author contributions

ZW: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review and editing. NK: Conceptualization, Investigation, Writing – original draft. LB: Conceptualization, Writing – review and editing. RB: Conceptualization, Methodology, Supervision, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgements

The authors would like to thank Qifan Deng for his contributions to the early stages of this work.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor M.X declared a past co-authorship with the author R.B.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/friot.2026.1883673/full#supplementary-material

References

Bai, H., Islam, M. T., Xu, M., and Buyya, R. (2026). Oracl: optimized reasoning for autoscaling via chain of thought with llms for microservices. *arXiv Preprint arXiv: 2602.05292.*

Bittencourt, L. F., Braghetto, K. R., Cordeiro, D., and Sakellariou, R. (2024). "On digital twins for cloud continuum applications," in *International Conference on the Economics of Grids, Clouds, Systems, and Services* (Springer), 286–293.

Bittencourt, L. F., Rodrigues-Filho, R., Spillner, J., De Turck, F., Santos, J., da Fonseca, N. L., et al. (2025). The computing continuum: past, present, and future. *Comput. Sci. Rev.* 58, 100782. doi:10.1016/j.cosrev.2025.100782

Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, 13–16.

Cao, L., Huo, T., Li, S., Zhang, X., Chen, Y., Lin, G., et al. (2024). Cost optimization in edge computing: a survey. *Artif. Intell. Rev.* 57, 312. doi:10.1007/s10462-024-10947-4

Chen, Y., Sun, Y., Yu, H., and Taleb, T. (2024). Joint task and computing resource allocation in distributed edge computing systems via multi-agent deep reinforcement learning. *IEEE Trans. Netw. Sci. Eng.* 11, 3479–3494. doi:10.1109/TNSE.2024.3375374

Fan, W., Zhang, Y., Zhou, G., and Liu, Y. (2024). Deep reinforcement learning-based task offloading for vehicular edge computing with flexible RSU-RSU cooperation. *IEEE Trans. Intelligent Transp. Syst.* 25, 7712–7725. doi:10.1109/TITS.2024.3349546

Goudarzi, M., Palaniswami, M., and Buyya, R. (2022). Scheduling IoT applications in edge and fog computing environments: a taxonomy and future directions. *ACM Comput. Surv.* 55, 1–41. doi:10.1145/3544836

Jovičević-Klug, M., Souza Filho, I. R., Springer, H., Adam, C., and Raabe, D. (2024). Green steel from red mud through climate-neutral hydrogen plasma reduction. *Nature* 625, 703–709. doi:10.1038/s41586-023-06901-z

Kulandaivel, S., and Akuthota, U. C. (2026). Driving the future with edge computing: a review of data offloading, enabling technologies and trends for autonomous vehicles. *Internet Things* 38, 101963. doi:10.1016/j.iot.2026.101963

Li, Y., Zhang, X., Zeng, T., Duan, J., Wu, C., Wu, D., et al. (2023). Task placement and resource allocation for edge machine learning: a GNN-Based multi-agent reinforcement learning paradigm. *IEEE Trans. Parallel Distributed Syst.* 34, 3073–3089. doi:10.1109/TPDS.2023.3313779

Liang, Y., Xu, C., Mei, Y., Zou, H., Guo, J., Fan, X., et al. (2026). A comprehensive survey on large language model compression for artificial intelligence applications in edge systems. *IEEE Internet Things J.* 13, 20583–20599. doi:10.1109/JIOT.2026.3675866

Ma, S., Chai, Z., Li, Y., Fu, J., and Sun, L. (2025). Computation offloading for vehicular networks via sparse-attention-based multi-agent reinforcement learning. *Ad Hoc Netw.* 179, 104006. doi:10.1016/j.adhoc.2025.104006

Qiu, B., Wang, Y., Xiao, H., and Zhang, Z. (2024). Deep reinforcement learning-based adaptive computation offloading and power allocation in vehicular edge computing networks. *IEEE Trans. Intelligent Transp. Syst.* 25, 13339–13349. doi:10.1109/TITS.2024.3391831

Rasouli, N., Klein, C., and Elmroth, E. (2026). Resource management for mission-critical applications in edge computing: systematic review on recent research and open issues. *ACM Comput. Surv.* 58, 1–37. doi:10.1145/3762181

Satyanarayanan, M., Harkes, J., Blakley, J., Meunier, M., Mohandoss, G., Friedt, K., et al. (2022). Sinfonia: cross-Tier orchestration for edge-native applications. *Front. Internet Things* 1, 1025247. doi:10.3389/friot.2022.1025247

Sethi, S. P., Das, D. P., and Behera, S. K. (2023a). Monitoring of arc plasma process parameter using cnn-based deep learning algorithm to accommodate sensor failure. *IEEE Trans. Plasma Sci.* 51, 1434–1445. doi:10.1109/tps.2023.3274788

Sethi, S. P., Das, D. P., Behera, S. K., and Ray, N. (2023b). Instability and fault analysis of arc plasma using advanced signal processing methods. *Rev. Sci. Instrum.* 94, 104710. doi:10.1063/5.0161187

Srirama, S. N. (2024). A decade of research in fog computing: relevance, challenges, and future directions. *Softw. Pract. Exp.* 54, 3–23. doi:10.1002/spe.3243

Tong, L., Zhang, X., and Yan, Y. (2026). "Cloud-edge-end coordinated meta-reinforcement learning for adaptive resource allocation in intelligent transportation systems," in Algorithms and Architectures for Parallel Processing: 25th International Conference, ICA3PP 2025 (Springer), 175–186. doi:10.1007/978-981-95-8420-8_14

Transforma Insights (2025). Global IoT forecast report, 2024–2034.Available online at: https://transformainsights.com/research/reports/global-iot-forecast-report-2024-2034.

Vali, A. A., Azizi, S., Shojafar, M., and Buyya, R. (2026). Energy-efficient resource management in microservices-based fog and edge computing: state-Of-the-Art and future directions. *ACM Comput. Surv.* 58, 1–42. doi:10.1145/3797911

Vetriveeran, D., and Leena Sri, R. (2025). Resource provisioning in fog computing: a survey. *ACM Comput. Surv.* 57, 1–26. doi:10.1145/3744662

Wang, R., Gao, Z., Zhang, L., Yue, S., and Gao, Z. (2025a). Empowering large language models to edge intelligence: a survey of edge efficient LLMs and techniques. *Comput. Sci. Rev.* 57, 100755. doi:10.1016/j.cosrev.2025.100755

Wang, Z., Goudarzi, M., and Buyya, R. (2025b). ReinFog: a deep reinforcement learning empowered framework for resource management in edge and cloud computing environments. *J. Netw. Comput. Appl.* 242, 104250. doi:10.1016/j.jnca.2025.104250

Wang, Z., Goudarzi, M., and Buyya, R. (2025c). TF-DDRL: a transformer-enhanced distributed DRL technique for scheduling IoT applications in edge and cloud computing environments. *IEEE Trans. Serv. Comput.* 18, 1039–1053. doi:10.1109/TSC.2025.3528346

Wang, Z., Goudarzi, M., Gong, M., and Buyya, R. (2026). A knowledge distillation-empowered adaptive federated reinforcement learning framework for multi-domain IoT applications scheduling. *IEEE Trans. Mob. Comput.* 1–17. doi:10.1109/TMC.2026.3672603

Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., and Chen, J. (2025). A review on edge large language models: design, execution, and applications. *ACM Comput. Surv.* 57, 1–35. doi:10.1145/3719664

Zhou, G., Tian, W., Buyya, R., Xue, R., and Song, L. (2024). Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions. *Artif. Intell. Rev.* 57, 124. doi:10.1007/s10462-024-10756-9

Zhu, X., Li, J., Liu, Y., Ma, C., and Wang, W. (2024). A survey on model compression for large language models. *Trans. Assoc. Comput. Linguistics* 12, 1556–1577. doi:10.1162/tacl_a_00704