International Journal of Grid and High Performance Computing
Volume 18 • Issue 1 • January-December 2026

AI-Driven Multi-Agent Real-Time Load Balancing for Energy-Efficient Cloud–Edge Systems

Jianfeng Chen
Powerchina Huadong Engineering Corporation, China

Sijing Zhu
http://orcid.org/0009-0001-1846-4984
Powerchina Huadong Engineering Corporation, China

Anke Li
http://orcid.org/0009-0004-3158-1183
Powerchina Huadong Engineering Corporation, China

Yi Xue
Powerchina Huadong Engineering Corporation, China
Received: November 14th, 2025 | **Accepted:** March 26th, 2026

ABSTRACT

This study presents AI-driven real-time load balancing (AI-RTLB), a framework for heterogeneous cloud–edge–Internet of Things (IoT) systems. AI-RTLB combines long short-term memory attention-based workload forecasting with adaptive multi-agent reinforcement learning (decentralized actors, shared critic) and a post-optimization layer that enforces energy, fairness, thermal, and power-budget constraints. This design anticipates demand surges, coordinates distributed decisions, and produces service-level agreement (SLA)-aware schedules. Across large data-center traces, edge workloads, and IoT demand logs, AI-RTLB reduces average latency by 16.3% and improves energy efficiency by 21.7% over strong baselines, while increasing throughput and lowering SLA violations (5.3%) with a lower fairness index (0.041). Convergence is faster and more stable than with single-agent deep reinforcement learning, and robustness is maintained under workload noise and mixed job types. An ablation study confirms complementary gains from prediction, multi-agent control, and fairness- and energy-aware optimization. AI-RTLB offers a practical path toward efficient, equitable, and sustainable computing.

KEYWORDS

Real-Time Load Balancing, Multi-Agent Reinforcement Learning, Energy-Aware Optimization, Fairness-Aware Scheduling, Sustainable Computing

INTRODUCTION

The rapid proliferation of cloud computing, edge intelligence, and large-scale data centers has significantly increased the demand for efficient resource management. Among these challenges, real-time load balancing and energy efficiency optimization are critical to sustaining high-performance, large-scale distributed computing systems while mitigating environmental impact (Dahule, 2024). The International Energy Agency (2024) reports that data centers account for nearly 1–1.5% of global

DOI: 10.4018/IJGHPC.411217

electricity consumption, a figure projected to rise substantially with the growing adoption of AI-driven services and 6G networks. In this context, integrating AI into system-level scheduling and resource management has become an important direction for ensuring both computational sustainability and operational reliability. Achieving real-time optimization is increasingly important from both technical and ecological perspectives (Li, 2024).

Despite extensive research in load balancing and energy-aware scheduling, existing methods often suffer from critical limitations when applied to large-scale and heterogeneous distributed systems. Traditional rule-based approaches lack adaptability to dynamic workloads, while heuristic algorithms provide suboptimal solutions under stringent latency constraints (Rasoulnia, 2025). Similarly, classical machine learning methods struggle with generalization across heterogeneous computing environments and fail to capture temporal fluctuations in user demand (Wu et al., 2024). Recent deep reinforcement learning techniques have shown promise but face challenges in convergence speed, stability, and interpretability, which restrict their deployment in production-level scheduling systems (Shukla et al., 2025). Furthermore, many state-of-the-art methods prioritize either load balancing or energy efficiency in isolation, neglecting the interdependent trade-offs that characterize real-world computing infrastructures (Aldossary et al., 2025).

To address these gaps, this study proposes an AI-driven real-time optimization framework that integrates predictive modeling, adaptive control, and fairness-aware decision making, with a primary focus on system-level scheduling performance rather than algorithmic novelty. The first innovation is predictive workload modeling, which leverages temporal-spatial sequence learning to anticipate traffic surges and proactively allocate resources, offering proactive allocation compared with static-threshold or short-term averaging methods commonly used in prior work. The second innovation is adaptive multi-agent control, where a distributed reinforcement learning paradigm coordinates multiple servers or edge nodes to achieve global equilibrium under local constraints, improving scalability relative to centralized controllers commonly adopted in grid and cluster scheduling. The third innovation is energy-aware optimization, which incorporates power consumption models and thermal dynamics into the objective function to jointly minimize energy overhead and latency, addressing limitations of approaches that focus solely on performance metrics. Finally, the framework introduces fairness-constrained scheduling, ensuring equitable resource distribution across tasks and preventing starvation in multi-tenant environments, a dimension often neglected in prior scheduling systems. Together, these components form an integrated scheduling framework tailored for heterogeneous cloud–edge infrastructures, rather than a standalone machine learning model.

Empirical validation demonstrates the effectiveness of the proposed approach. Across three benchmark datasets derived from real-world data center traces, the framework achieves a 16.3% reduction in average latency, a 21.7% improvement in energy efficiency, and a 12.5% increase in throughput compared with strong baselines such as round-robin (RR), power-aware load balancer, and state-of-the-art deep reinforcement learning schedulers. Statistical significance tests confirm that the improvements are consistent ($p < .01$), while convergence analysis shows accelerated training stability relative to conventional methods. Beyond numerical performance, the proposed framework has potential applications in large-scale distributed scheduling scenarios, including cloud service platforms, edge computing networks, and AI inference clusters. From an academic perspective, it contributes to the literature by bridging AI-assisted decision making with classical distributed scheduling objectives, providing a replicable experimental framework and benchmark results for sustainable computing research.

RELATED WORKS

Application Scenarios and Challenges

Real-time load balancing and energy efficiency optimization are central to the operation of large-scale cloud data centers, edge computing platforms, and emerging Internet of Things (IoT)