Gyan Manthan
Journal homepage: www.gyanmanthan.org ISSN **3108-3455**

A Comprehensive Review of AI Techniques for Dynamic Resource Scheduling in Cloud Computing

Mahesh Patel
Assistant Professor, Dept. of Computer Science and Application, Arihant College, Indore

Abstract:

This review paper focuses on the challenge of dynamic resource scheduling in cloud computing environments, which is complicated by unpredictable workloads, variable demand, and strict Quality of Service (QoS) requirements. Traditional scheduling methods struggle to meet these demands in heterogeneous and rapidly changing cloud settings. To address this, the study explores the application of Artificial Intelligence (AI) techniques—including machine learning (ML), deep learning (DL), reinforcement learning (RL), and hybrid models—for intelligent and adaptive resource management.

The paper systematically categorizes these AI-based approaches based on their core methodologies, decision-making criteria (e.g., cost, energy, performance), and their ability to adapt in real-time. It also provides a comparative analysis highlighting each method's strengths and limitations in terms of scalability, prediction accuracy, convergence speed, and computational overhead.

Key findings indicate that while AI enhances decision-making and efficiency in resource scheduling, challenges remain in areas such as SLA-aware orchestration, energy efficiency, and integrating edge-cloud optimization. The paper emphasizes the need for future research in cross-layer optimization, explainable AI (XAI), and real-world implementation benchmarks. These insights aim to guide future developments in building more robust, intelligent, and practical AI-driven cloud resource management systems.

Keywords: Resource Scheduling, Machine learning, Deep Learning, Cloud computing.

Introduction:

With its scalable, adaptable, and reasonably priced infrastructure for hosting apps and services, cloud computing has completely changed how computing resources are provided and used. However, effectively managing resources gets more difficult as the demand for cloud services rises. Allocating virtualized resources (such as CPU, memory, and storage) in real-time based on changing workloads, service-level agreements (SLAs), and energy consumption limitations is one of the most difficult problems.

These dynamic requirements are frequently not met by conventional static and rule-based scheduling techniques, which can result in problems like underutilization, overprovisioning, SLA violations, and higher operating expenses. Artificial Intelligence (AI) approaches have therefore surfaced as viable means of automating and optimizing resource scheduling. AI models are able to make wise decisions in real time, adjust to shifting patterns, and learn from past data.

An extensive analysis of AI-based methods for cloud computing's dynamic resource scheduling is provided in this paper. It investigates how to optimize scheduling choices using machine learning (ML), deep learning (DL), reinforcement learning (RL), and hybrid artificial intelligence (AI) models. We classify current research according to algorithmic approaches, optimization goals like performance, cost-effectiveness, energy consumption, and QoS compliance, and decision parameters. The review also points out the shortcomings of the models that are currently in use and suggests some possible avenues for further study, such as explainable AI (XAI), energy-aware scheduling, and cloud-edge resource integration. For researchers and practitioners looking to create intelligent, flexible, and scalable cloud resource management systems, this study intends to act as a fundamental resource.

Objectives of the Study

The purpose of this research is to methodically investigate and assess how artificial intelligence (AI) functions in dynamic resource scheduling in cloud computing settings. The particular goals are:

1. To evaluate the shortcomings of conventional cloud resource management strategies in managing diverse environments and dynamic workloads.
2. To examine and classify AI-based methods for scheduling cloud resources, such as machine learning (ML), deep learning (DL), reinforcement learning (RL), and hybrid models.
3. To evaluate AI methods according to important performance indicators like computational overhead, energy efficiency, convergence speed, prediction accuracy, and scalability.
4. To determine research gaps in the current approaches to AI-driven resource management, specifically in areas such as system generalizability, edge-cloud integration, energy optimization, and SLA-awareness.
5. To make recommendations for future research areas, such as the use of federated learning, explainable AI (XAI), and cross-layer optimization for safe and useful cloud resource management.

These goals direct the paper's contribution to the development of AI-powered cloud infrastructure that is more intelligent, flexible, and effective.

Literature Review

Over time, cloud resource management has evolved from rule-based scheduling to intelligent, adaptive systems powered by Artificial Intelligence (AI). Early work by **Buyya et al. (2020)** emphasized the benefits of data-driven resource management in heterogeneous, multi-tier environments, demonstrating improved flexibility and energy-efficient GPU scaling using AI on platforms like Google Cloud and Azure.

In **2021**, **Venkateswarlu et al.** introduced a hybrid framework combining Machine Learning (ML) and Reinforcement Learning (RL) for automated resource provisioning and demand prediction. Their model enhanced scalability and cost-efficiency, laying groundwork for intelligent scheduling systems.

Saad Iqbal and Ann Heng (2023) advanced this by integrating AI with edge computing and IoT, using real-time sensor data to support proactive, context-aware scheduling through ML models.

In **2024**, **Nandyala et al.** demonstrated the superior performance of RL and supervised learning in multi-cloud scenarios, achieving up to 90% resource efficiency and 25% cost savings during peak demand. **Annam (2024)** focused on DL techniques like LSTM and Random Forests for improved anomaly detection and demand forecasting, showcasing real-time scalability and operational savings.

Most recently, **Wang and Xing (2025)** targeted CPU resource optimization in cloud operating systems, illustrating RL's effectiveness in maximizing utilization and reducing manual intervention.

Despite progress, gaps remain in platform generalization, multi-cloud and edge integration, and explainability—highlighting future research directions in XAI, federated learning, and hybrid AI models.

AI Techniques and Their Effectiveness in Cloud Resource Management

Sr. No.	AI Technique	Application Area	Key Functions	Effectiveness / Outcome
1	**Machine Learning (ML)**	Resource prediction, anomaly detection	Forecasting CPU, memory usage; identifying abnormal usage patterns	Improved workload prediction accuracy; reduced manual configuration
2	**Deep Learning (DL)**	Complex workload forecasting, auto-scaling	Time-series prediction using LSTM, CNN, RNN	20–30% better accuracy in workload prediction vs. traditional models
3	**Reinforcement Learning (RL)**	Dynamic resource scheduling, VM migration	Real-time decisions based on rewards; learning from environment feedback	Up to 25% cost savings; 90% resource efficiency during peak demand
4	**Hybrid Models (e.g. RL+GA)**	Multi-objective optimization	Combines exploration (GA) with adaptive learning (RL)	Balanced resource utilization; scalable in heterogeneous workloads

5	**Bayesian Optimization**	Task scheduling	Probabilistic modeling of workload distributions	Reduced job completion time and improved scheduler efficiency
6	**Fuzzy Logic**	SLA-aware resource reservation	Decision-making under uncertainty	Enhanced SLA compliance and user satisfaction
7	**Swarm Intelligence (ACO/PSO)**	Load balancing, service selection	Bio-inspired optimization (ants, particles)	High energy efficiency; optimized response time
8	**Supervised Learning**	Predictive auto-scaling	Uses labeled historical data for model training	Up to 92% accuracy (e.g., Random Forest) in predicting workload demands
9	**Unsupervised Learning**	Anomaly detection, usage pattern recognition	Identifies patterns in unlabeled data	Effective in detecting unusual behavior and outliers
10	**Q-Learning (a type of RL)**	Autoscaling, resource migration	Learns optimal policy through rewards and penalties	Improved system reliability and reduced latency

Challenges and Limitations

To guarantee dependable and effective performance, AI-driven resource scheduling in cloud environments must overcome a number of important obstacles. Computational overhead and model complexity are two main issues. Complex artificial intelligence models, like deep learning and hybrid reinforcement learning, frequently call for large amounts of memory and processing power, which can cause latency and performance snags, particularly in real-time cloud operations. Furthermore, because most AI models are trained on static datasets, they are unable to adapt in real-time to unexpected workload fluctuations or anomalies. Even though online education presents a viable remedy, its use in expansive, dynamic cloud environments is still in its infancy.

Critical problems also arise with data availability and quality. Both historical and real-time monitoring data are crucial for AI models, and the presence of noisy, biased, or incomplete data can lead to poor scheduling choices, erroneous predictions, or even Service Level Agreement (SLA) violations. The implementation of AI in cloud scheduling is made more difficult by ethical, security, and privacy issues. While federated learning and privacy-preserving techniques show promise, they are not yet widely used in practice. Inadequately governed AI systems can unintentionally introduce bias or leak sensitive information.

Differences in APIs, SLAs, and configuration standards amongst providers, including AWS, Azure, and Google Cloud, make resource scheduling challenging. Many AI models still lack the interoperability required for cross-platform operations because they were created for

single-cloud environments. Furthermore, the limited scalability and generalization of AI models prevent their wider application. Scaling models to support large, distributed, and hybrid systems is still an open research question because models trained on particular workloads or cloud configurations may not function well in other or quickly changing environments. Finally, a recurring obstacle is the lack of explainability and confidence in AI judgments. Many AI systems operate as "black boxes," which makes it challenging for administrators to debug, understand, or trust them.

Conclusion:

Dynamic resource scheduling is fundamental to ensuring the performance, scalability, and cost-efficiency of modern cloud computing environments. Traditional scheduling techniques, while effective in simpler contexts, struggle to cope with the complexity and variability of today's cloud workloads. This review has highlighted how Artificial Intelligence (AI), particularly techniques such as Machine Learning (ML), Deep Learning (DL), Reinforcement Learning (RL), and hybrid models, offers a transformative approach to resource management.

AI-driven scheduling models have demonstrated substantial improvements in areas such as workload prediction, automated scaling, SLA compliance, and energy optimization. Reinforcement Learning and its variants stand out for their adaptability and decision-making in real-time environments, while deep learning excels at capturing complex workload patterns. However, alongside these advancements, there are also notable challenges—ranging from high computational costs and data privacy concerns to limited explainability and cross-cloud interoperability.

Despite these limitations, the potential of AI in cloud resource scheduling is immense. The ongoing evolution toward edge–cloud synergy, green computing, and federated AI opens new avenues for research and innovation. Addressing the existing challenges through explainable AI, privacy-preserving models, and standard benchmarking frameworks will be crucial to realizing AI's full potential in this domain.

In conclusion, AI is not just a supplementary tool but a strategic enabler for intelligent and autonomous cloud resource management. Future research should focus on building scalable, secure, and adaptive AI systems that align with the dynamic and distributed nature of next-generation cloud infrastructures.

Recommendations and Future Directions

As AI continues to revolutionize cloud computing, integrating advanced AI techniques into dynamic resource scheduling presents promising opportunities. However, several limitations and unexplored areas must be addressed to realize its full potential. Based on the literature review, the future research directions are proposed as to enhance the effectiveness of AI-driven resource management in cloud environments, several future directions can be considered. One promising approach is the adoption of hybrid AI models by integrating Reinforcement Learning (RL) with other techniques such as Genetic Algorithms (GA), Swarm Intelligence, or Deep Learning. This

combination can significantly improve the adaptability and performance of scheduling mechanisms, especially in complex, multi-objective scenarios. Additionally, there is a growing need to prioritize Service Level Agreement (SLA) and Quality of Service (QoS) awareness. Future scheduling solutions should explicitly incorporate SLA parameters to ensure service reliability and regulatory compliance. Another critical area is the promotion of real-time learning systems. By utilizing online learning techniques and continuous feedback loops, systems can dynamically adapt to changing workloads and operating conditions. Furthermore, the development of interoperable frameworks is essential, enabling standardized AI models that function seamlessly across multiple cloud service providers in a multi-cloud setup. Lastly, it is vital to ensure transparency and interpretability in AI decision-making. The integration of Explainable AI (XAI) frameworks will help make scheduling decisions more understandable, which is especially important for mission-critical and regulated applications.

Future advancements in AI-driven cloud resource management should focus on several strategic areas to enhance efficiency, scalability, and ethical compliance. One key direction is the integration of edge and cloud resources, where AI models can dynamically coordinate between cloud and edge environments to support latency-sensitive applications such as IoT, augmented/virtual reality (AR/VR), and autonomous systems. Another critical focus is the development of energy-efficient and green AI scheduling techniques that reduce carbon footprints while maintaining system performance through energy-aware resource allocation. Privacy and data security are also gaining importance, leading to the adoption of federated and privacy-preserving learning approaches. These enable AI models to be trained across distributed cloud nodes without centralizing sensitive user data, ensuring compliance with data protection regulations. Furthermore, AI governance and ethical design must be prioritized to address issues of bias, fairness, and transparency in scheduling decisions, especially in public or shared cloud infrastructures. The integration of quantum computing with AI presents a promising frontier, potentially accelerating complex optimization tasks in large-scale cloud resource scheduling. Finally, there is a strong need for standardized benchmarking and validation frameworks. Establishing realistic simulation environments and utilizing open datasets, such as CloudSim or Google Cluster Data, will help evaluate and validate AI models effectively under real-world cloud conditions.

References:

[1] C. H. Venkateswarlu, K. Chiranjeevi, and V. Kumar, “AI-POWERED CLOUD RESOURCE MANAGEMENT: ENHANCING EFFICIENCY, SCALABILITY, AND COST OPTIMIZATION.” [Online].

[2] A. Kumar, “AI-Driven Innovations in Modern Cloud Computing,” *Computer Science and Engineering*, vol. 2024, no. 6, pp. 129–134, doi: 10.5923/j.computer.20241406.02.

[3] T. Resurssien and A. Ja, “AI DRIVEN OPTIMIZATION OF RESOURCE ALLOCATION
AND COST EFFICIENCY IN CLOUD COMPUTING ENVIRONMENTS.”

[4] Y. Wang and S. Xing, “AI-Driven CPU Resource Management in Cloud Operating Systems,” *Journal of Computer and Communications*, vol. 13, no. 06, pp. 135–149, 2025, doi: 10.4236/jcc.2025.136009.

[5] S. Banerjee, "Intelligent Cloud Systems: AI-Driven Enhancements in Scalability and Predictive Resource Management." Jan. 16, 2025. doi: 10.20944/preprints202501.1153.v1.

[6] N. Annam, "AI-Driven Solutions for IT Resource Management," *International Journal of Engineering and Management Research*, vol. 14, no. 6, pp. 15–30, Dec. 2024, doi: 10.31033/ijemr.14.6.15-30.

[7] Somnath Banerjee, "Intelligent Cloud Systems: AI-Driven Enhancements in Scalability and Predictive Resource Management," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 266–276, Dec. 2024, doi: 10.48175/ijarsct-22840.

[8] C. Lekkala, "Citation: Lekkala C. AI-Driven Dynamic Resource Allocation in Cloud Computing: Predictive Models and Real-Time Optimization," *J Artif Intell Mach Learn & Data Sci*, vol. 2024, no. 2, pp. 450–456, 2024, doi: 10.51219/JAIMLD/chandrakanth.

[9]A. K. Nandyala, M. P. Gore, and N. Gupta, "Using AI To Optimize Resource Allocation In Multi-Cloud Environments," 2024. [Online]. Available: www.ijcrt.org

[10] R. Kedarnath Navandar, K. Naik, M. K. Patil, P. P. Kothari, and D. Desai, "Enhancing Cloud Computing Environments with AI-Driven Resource Allocation Models," 2024.

[11] A. Heng and S. Iqbal, "AI-Driven Resource Management in Cloud Computing: Leveraging Machine Learning, IoT Devices, and Edge-to-Cloud Intelligence," 2023, doi: 10.13140/RG.2.2.28383.27049.

[12] Prof. Dr. A. S. Rao, "Orchestrating Efficiency: AI-Driven Cloud Resource Optimization for Enhanced Performance and Cost Reduction," *International Journal of Research Publication and Reviews*, vol. 4, no. 12, pp. 2007–2009, Dec. 2023, doi: 10.55248/gengpi.4.1223.123430.

[13] S. Chinamanagonda, "AI-Driven Cloud Cost Management - AI Tools For Optimizing Cloud Resource Allocation and Costs," *International Journal of Science and Research (IJSR)*, vol. 12, no. 12, pp. 2135–2149, Dec. 2023, doi: 10.21275/sr24829170724.

[14] S. Ilager, R. Muralidhar, and R. Buyya, "Artificial Intelligence (AI)-Centric Management of Resources in Modern Distributed Computing Systems," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2006.05075

[15] Md. S. Hasan, "Artificial Intelligence (AI) Backed Cloud Resource Management Approach for Infrastructure as a Service (IAAS)," *TEXILA INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH*, pp. 80–86, Dec. 2019, doi: 10.21522/tijar.2014.se.19.02.art010.