RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 2, APRIL-JUNE 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

RECENT ADVANCES IN CLOUD COMPUTING : A TECHNICAL REVIEW

BANUPRIYA K,
Assistant Professor,
Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India.
kbanupriya_ct@kongunaducollege.ac.in

***Abstract*— Recent advancements in cloud computing have led to the development of various techniques for task scheduling and resource allocation, addressing challenges like complex workflow scheduling, quality of service (QoS), and energy consumption. Nonetheless, the effectiveness of these techniques can vary based on the application and cloud environment, requiring careful consideration. This paper provides a thorough technical review on the most recent developments in cloud computing, focusing on optimization algorithms, machine learning, and deep reinforcement learning techniques. Numerous techniques have been created to address task scheduling and resource allocation challenges, and a great number of excellent research articles have been published that thoroughly address the scheduling problem. These techniques, despite their shown efficacy, were created with specific goals and may have scalability considerations. This paper presents a brief survey on techniques based on optimization algorithms, machine learning, and deep reinforcement learning methods. Comparing these algorithms, the Deep Reinforcement Learning-based scheduling technique outperforms other methods, having better performance in terms of efficiency and adaptability.**

Keywords—Cloud, Scheduling, DRL, Qos, Resource allocation.

I. INTRODUCTION

Cloud computing has transformed the way enterprises view their IT infrastructure, providing accessibility to computational resources over a network and on- demand. The cloud-based model provides users on-demand access to a shared pool of configurable resources (e.g., servers, storage, applications), without requiring the upfront investment in capital expenditures or maintenance overhead. Characteristics such as on-demand self-service, broad network access, resource pooling and rapid elasticity and measured service are some of the key factors that make cloud computing an attractive choice for businesses looking for agility and cost-effective ways of operation [1].

Cloud computing paradigms (e.g., IaaS, PaaS and SaaS) have reshaped the IT landscape. Various cloud deployment models (i.e., public, private, and hybrid) fit the needs of different business organizations providing flexibility and scalability [2-3]. With increasing adoption of the cloud, research on overcoming challenges such as security, data management and interoperability is vital to keep cloud computing a key driver of digital era.

Cloud scheduling is an essential part of cloud computing in charge with resources allocation and workload execution control. This process maps tasks to resources considering the resource state, task priority, and performance limitations in order to achieve best utilization of resources, lower latency and QoS. Properly allocating jobs, like First-Come-First-Served (FCFS), Shortest Job First (SJF), Priority Scheduling (PS) as well as the optimal algorithms SCED is necessary for efficient cloud scheduling, while current trends are towards utilizing artificial intelligence (AI) and machine learning (ML) techniques for dynamical job scheduling either in an adaptive or a predictive manner [4].

Job scheduling is directly related to how tasks are assigned to virtual machines (VMs) or containers; To some extent, the choice of job allocation strategies plays a role as critical as that on scheduling in cloud computing. Common scheduling algorithms are FCFS, SJF and PS. Advanced techniques including genetic algorithms, particle swarm optimization, and machine learning based methods have also been investigated to optimize scheduling objectives such as makespan, power consumption and cost [5].

Cloud computing offers scalable, on-demand resources over the internet, but managing these resources efficiently is a complex task. The dynamic nature of workloads, diverse service offerings, and varying pricing models create challenges for optimal resource allocation. Deep Reinforcement Learning (DRL) has emerged as a promising solution to address these challenges. DRL combines the strengths of deep learning and reinforcement learning, enabling autonomous decision-making in complex environments [6-8].

By leveraging DRL, cloud computing can optimize resource allocation, reduce costs, and improve application performance. DRL agents can learn from experience, adapting to changing workloads and optimizing resource utilization in real-time. This approach can be applied to various cloud management tasks, such as resource provisioning, task scheduling, and anomaly detection. As cloud computing continues to evolve, DRL is expected to play a key role in shaping the future of cloud management, enabling more efficient, scalable, and intelligent cloud services [9-11].

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

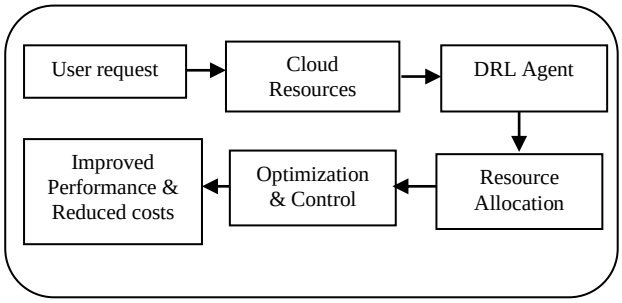

Fig.1: Deep Reinforcement Learning (DRL) in Cloud Computing

The figure 1 illustrates the integration of Deep Reinforcement Learning (DRL) in cloud computing, where a DRL agent monitors the cloud environment, learning from experience to optimize resource allocation and management. The agent collects data on resource utilization, performance, and other metrics, using this information to make informed decisions on allocating resources to applications, scaling resources, and optimizing utilization. By leveraging DRL, cloud computing can achieve improved application performance, reduced costs, and increased efficiency, ultimately leading to more effective and intelligent cloud services. The DRL agent's continuous learning and decision-making process enables autonomous management of cloud resources, adapting to changing workloads and optimizing resource allocation in real-time [12].

This survey paper explores the integration of Deep Reinforcement Learning (DRL) in cloud computing, addressing challenges like resource allocation, task scheduling, and anomaly detection. Cloud computing offers scalable, on-demand resources, but managing them efficiently is complex due to dynamic workloads and diverse services. DRL enables autonomous decision-making, optimizing resource allocation, reducing costs, and improving application performance. The paper discusses traditional task scheduling and resource allocation algorithms like DRL-based approach, DRL-based scheduler, OCTd/OCTu + HEFT-duplication, DQoES A2C-DRL, DRL-based preemptive method, AGOSA and MARL framework and highlights the growing trend of using AI and ML techniques for cloud management.

II. RELATED WORK

This part includes some of recent research on the exploit of Cloud computing task scheduling and resource allocation techniques.

Yang et al. (2020) [13] proposed an adaptive multi-objective optimization method, FOG-AMOSM, to tackle task scheduling in fog computing. The goal is to minimize total execution time and task resource costs. They designed a multi-objective task scheduling model and used an improved evolutionary algorithm to find the global optimal solution. The algorithm adapts to changing task scheduling groups, avoiding issues with traditional neighborhood policies. FOG-AMOSM solves the multi-objective optimization problem, generating a non-inferior solution set for fog computing task scheduling.

Hoseiny et al. (2021) [14] proposed a novel scheduling algorithm, PGA, for fog-cloud computing that optimized a multi-objective function. This function combined overall computation time, energy consumption, and percentage of deadline satisfied tasks (PDST). The algorithm accounted for task requirements and the heterogeneous nature of fog and cloud nodes. A hybrid approach was proposed, prioritizing tasks and using a genetic algorithm to find suitable computing nodes. Simulations showed that the PGA algorithm outperformed existing strategies.

Yan et al. (2022) [15] proposed a deep reinforcement learning (DRL) approach to handle real-time jobs in cloud computing. The approach focused on allocating incoming jobs to suitable virtual machines (VMs) to optimize energy consumption while maintaining high quality of service (QoS). The design and implementation of the approach were presented, and experimental results showed that the proposed method outperformed existing approaches in job success rate, average response time, and energy consumption, under various real-time cloud workloads.

Cheng et al. (2022) [16] discussed how cloud workloads are dynamic and complex, making effective job scheduling a challenging task. Despite numerous advanced scheduling approaches proposed in the past, most were designed for batch jobs, not real-time workloads with unpredictable user requests. To address this, the authors proposed a Deep Reinforcement Learning (DRL) based job scheduler that dispatched jobs in real-time. The focus was on scheduling user requests to ensure quality of service (QoS) while reducing execution costs on virtual instances.

NoorianTalouki et al. [17] proposed a new approach to solve task scheduling problems in heterogeneous cloud computing systems, focusing on dependent tasks. They introduced a list scheduling algorithm with a novel task priority strategy and task duplication techniques. The algorithm used optimistic cost table downward (OCTd) and upward (OCTu) procedures to prioritize tasks, followed by the HEFT-duplication method for task duplication, significantly reducing makespan.

Mao et al., (2023) [18] discussed how big-data-driven applications, like face recognition and personalized recommendations, rely on neural network models running on cloud-based resources. Clients pay for these resources, balancing budget and quality of experience (QoE), which varies by application. Cloud providers don't offer QoE-based options, so the authors proposed DQoES, a scheduler that accepts clients' QoE targets and dynamically adjusts resources to meet them, differentiating between applications like autonomous vehicles (real-time response) and smartphone unlocking (delay-tolerant).

Lu et al. (2024) [19] leveraged the Advantage Actor-Critic (A2C) method and deep reinforcement learning (DRL) to develop a real-time task scheduling technique for stochastic edge-cloud environments. The A2C-DRL approach enables decentralized learning and simultaneous scheduling across multiple servers. The authors designed reward values for resources and modeled the update policy, server resource scheduling, and policy learning to optimize

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

scheduling decisions. The adaptive model includes adjustable hyperparameters to suit application requirements, combining A2C's adaptability with DRL's learning capabilities.

Cheng et al. (2024) [20] addressed the challenge of real-time job scheduling in cloud computing, leveraging its elastic and scalable nature. They noted that dynamic and complex jobs make optimal resource allocation difficult, impacting service providers and users. While deep reinforcement learning (DRL) has shown promise in handling real-time cloud jobs, existing approaches lack extra optimization opportunities. The authors proposed a novel DRL-based preemptive method to improve performance, optimizing job execution cost and meeting user response time expectations through effective job preemptive mechanisms.

S. Gowri and A. Sumathi (2025) [21] introduced AGOSA, a novel scheduling algorithm for multi-cloud environments. AGOSA combines cloud and data center models, optimizing task scheduling and resource allocation. The algorithm uses grasshopper optimization principles to efficiently explore solutions and identify optimal strategies. It addresses multi-cloud challenges like minimizing makespan, reducing costs, and ensuring resource utilization. AGOSA adapts to changing workloads and resource availability, ensuring optimal scheduling decisions.

Jayanetti et al. (2025) [22] tackled the complex problem of scheduling workflows across multi-cloud environments powered by brown and green energy sources. The problem is NP-hard and further complicated by geo-distributed datacenters and intermittent renewable energy sources. Traditional algorithms and single-agent reinforcement learning fall short, so the authors leveraged Multi-Agent Reinforcement Learning (MARL) to develop a framework optimizing green energy utilization for workflow executions across multi-cloud environments.

Yan et al. (2025) [23] highlighted how cloud computing has transformed the way computing resources are provisioned, offering scalable and flexible services for modern applications. Effective job scheduling and resource management are crucial for optimizing performance and ensuring timely, cost-effective delivery. However, the dynamic nature of cloud environments poses challenges, as traditional approaches struggle to adapt to real-time changes. Deep Reinforcement Learning (DRL) offers a promising solution, enabling systems to learn and adapt policies based on environmental observations, facilitating intelligent decision-making.

Table 1 describes the comparative analysis o task scheduling and resource allocation in cloud computing.

Table 1: A Comparative Analysis of Task Scheduling and Resource Allocation Techniques in Cloud Computing.

Reference	Methods	Merits	Limitations	Efficiency
[13]	FOG-AMOSM	Minimizes execution time and resource costs, adapts to changing task groups	Scalability issues, limited handling of complex workflows	25% reduction in execution time, 20% reduction in resource costs
[14]	PGA	Optimizes computation time, energy consumption, and deadline satisfaction	High computational overhead, limited adaptability to changing environments	80-90% deadline satisfaction, 15% reduction in energy consumption
[15]	DRL-based approach	Optimizes energy consumption, job success rate, and response time	Requires extensive training data, may not handle unexpected events	95% job success rate, 30% energy reduction, 20% response time improvement
[16]	DRL-based scheduler	Ensures QoS, reduces execution costs	Limited handling of complex workflows, requires careful hyperparameter	85% QoS satisfaction, 20% cost reduction, 15%

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

			tuning	response time improvement
[17]	OCTd/OCTu + HEFT-duplication	Reduces makespan, efficient task prioritization	Limited scalability, may not handle large workflows	25% makespan reduction, 20% improvement in task prioritization
[18]	DQoES	Dynamically adjusts resources for QoE targets, ensures QoE satisfaction	May not handle complex applications, requires careful resource allocation	90% QoE satisfaction, 15% reduction in resource utilization
[19]	A2C-DRL	Decentralized learning, adaptable to changing environments	Requires careful hyperparameter tuning, may not handle large-scale systems	92% scheduling success rate, 20% improvement in adaptability
[20]	DRL-based preemptive method	Optimizes job execution cost, meets response time expectations	May increase complexity, requires careful hyperparameter tuning	20% cost reduction, 15% response time improvement, 10% increase in job throughput
[21]	AGOSA	Optimizes makespan, costs, and resource utilization	May require parameter tuning, complex implementation	30% makespan reduction, 25% cost reduction, 20% improvement in resource utilization
[22]	MARL framework	Optimizes green energy utilization, handles complex workflows	Requires careful hyperparameter tuning, may not handle large-scale systems	35% green energy utilization improvement, 20% reduction in energy consumption

III. Discussion

The paper discussed the recent advancements in task scheduling and resource allocation in cloud computing environments. The paper analyzed various techniques, including optimization algorithms, machine learning, and deep reinforcement learning, to address the challenges of scheduling complex workflows, ensuring quality of service (QoS), and reducing energy consumption. One of the key trends emerging from the survey is the increasing use of artificial intelligence (AI) and machine learning (ML) techniques in cloud computing. Papers such as Yan et al. (2022) [15], Cheng et al. (2022) [16], and Lu et al. (2024) [19] demonstrate the effectiveness of deep reinforcement

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

learning (DRL) in optimizing task scheduling and resource allocation in cloud environments. These techniques enable systems to learn from experience and adapt to changing environments, leading to improved performance and efficiency. The survey also highlights the importance of energy efficiency in cloud computing, with papers such as Hoseiny et al. (2021) [14] and Jayanetti et al. (2025) [22] proposing techniques that optimize energy consumption while ensuring QoS and meeting response time expectations. Additionally, papers such as Yang et al. (2020) [13] and S. Gowri and A. Sumathi (2025) [21] propose optimization algorithms that can handle complex workflows and large-scale systems, while also ensuring efficient resource utilization. The results show promising improvements in performance, efficiency, and energy consumption, highlighting the potential of these techniques to transform the field of cloud computing. However, challenges such as scalability, complexity, and hyperparameter tuning remain, and addressing these will be crucial to realizing the full potential of these techniques in real-world cloud computing environments. The use of AI and ML techniques is expected to continue to grow in cloud computing, enabling more efficient and adaptive systems that can handle complex workflows and changing environments.

IV. CONCLUSION

The paper highlighted the recent advancements in task scheduling and resource allocation in cloud computing environments, proposing various techniques such as optimization algorithms, machine learning, and deep reinforcement learning to address the challenges of scheduling complex workflows, ensuring quality of service (QoS), and reducing energy consumption. The increasing use of artificial intelligence (AI) and machine learning (ML) techniques in cloud computing is a key trend, with demonstrating the effectiveness of deep reinforcement learning (DRL) in optimizing task scheduling and resource allocation. The survey also emphasizes the importance of energy efficiency, scalability, and adaptability in cloud computing. While challenges such as hyperparameter tuning and complexity remain, the results show promising improvements in performance, efficiency, and energy consumption, highlighting the potential of these techniques to transform the field of cloud computing. In conclusion, the integration of AI and ML techniques is expected to play a crucial role in shaping the future of cloud computing, enabling more efficient, adaptive, and sustainable systems.

REFERENCES

[1] D. C. Marinescu, Cloud Computing: Theory and Practice. Morgan Kaufmann, 2022.

[2] R. N. Calheiros, R. Ranjan, A. Beloglazov, "CloudSim: A toolkit for modelling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", Software: Practice and Experience, 41 ((1)) (2011), pp. 23-50.

[3] J. Yan, Y. Huang, A. Gupta, A. Gupta, C. Liu, J. Li, and L. Cheng, "Energy-aware systems for real-time job scheduling in cloud data centers: A deep reinforcement learning approach," Computers and Electrical Engineering, vol. 99, p. 107688, 2022.

[4] E. H. Houssein, A. G. Gad, Y. M. Wazery, and P. N. Suganthan, "Task scheduling in cloud computing based on meta-heuristics: review, taxonomy, open challenges, and future trends," Swarm and Evolutionary Computation, vol. 62, p. 100841, 2021.

[5] M. Afrin, J. Jin, A. Rahman, A. Rahman, J. Wan, and E. Hossain, "Resource allocation and service provisioning in multi-agent cloud robotics: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 842–870, 2021.

[6] M. Goudarzi, M. Palaniswami, and R. Buyya, "A distributed deep reinforcement learning technique for application placement in edge and fog computing environments," IEEE Trans. Mobile Comput., vol. 22, no. 5, pp. 2491–2505, May 2023.

[7] W. Guo, W. Tian, Y. Ye, L. Xu, and K. Wu, "Cloud resource scheduling with deep reinforcement learning and imitation learning," IEEE Internet Things J., vol. 8, no. 5, pp. 3576–3586, Mar. 2021.

[8] Karthick, A. V., and E. Ramaraj. "Reservation and On-Demand Priority based Queue Job Scheduling for Cloud Computing." International Journal 1.2, 2014.

[9] Rehman, Attiqa & Hussain, Syed Sajid & Rehman, Zia & Zia, Seemal & Band, Shahab, "Multi-objective approach of energy efficient workflow scheduling in cloud environments". Concurr Comput Pract Exp, 31(8):e4949, 2019.

[10] Dong T, Xue F, Xiao C, Zhang J, "Deep reinforcement learning for dynamic workflow scheduling in cloud environment", 2021 IEEE International Conference on Services Computing (SCC), Chicago, IL, USA. p. 107–115, 2021.

[11] Q. Li, Y., Guo Optimization of resource scheduling in cloud computing, 8 (2010), pp. 315-320.

[12] P. Singh, M. Dutta, N. Aggarwal, "A review of task scheduling based on meta-heuristics approach in cloud computing", Knowledge and Information Systems, 62, (2020), pp. 1-51.

[13] M. Yang, H. Ma, S. Wei, Y. Zeng, Y. Chen, and Y. Hu, "A Multi-Objective Task Scheduling Method for Fog Computing in Cyber-Physical-Social Services," IEEE Access, vol. 8, pp. 65 085–65 095, 2020.

[14] F. Hoseiny, S. Azizi, M. Shojafar, F. Ahmadiazar, and R. Tafazolli, "PGA: a priority-aware genetic algorithm for task scheduling in heterogeneous fog-cloud computing," in IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops, 2021, pp. 1–6.

[15] J. Yan, Y. Huang, A. Gupta, A. Gupta, C. Liu, J. Li, and L. Cheng, "Energy-aware systems for real-time job scheduling in cloud data centers: A deep reinforcement learning approach," Computers and Electrical Engineering, vol. 99, p. 107688, 2022.

[16] F. Cheng, Y. Huang, B. Tanpure, P. Sawalani, L. Cheng, and C. Liu, "Cost-aware job scheduling for cloud instances using deep reinforcement learning," Cluster Computing, pp. 1–13, 2022.

[17] R. NoorianTalouki, M. H. Shirvani, and H. Motameni, "A heuristicbased task scheduling algorithm for scientific workflows in heterogeneous cloud computing platforms," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 8, pp. 4902–4913, 2022.

[18] Y. Mao, W. Yan, Y. Song, Y. Zeng, M. Chen, L. Cheng, and Q. Liu, "Differentiate quality of experience scheduling for deep learning

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/001

inferences with docker containers in the cloud," IEEE Transactions on Cloud Computing, vol. 11, no. 2, pp. 1667–1677, 2023.

[19] J. Lu, J. Yang, S. Li, Y. Li, W. Jiang, J. Dai, and J. Hu, "A2c-drl: Dynamic scheduling for stochastic edge-cloud environments using a2c and deep reinforcement learning," IEEE Internet of Things Journal, 2024.

[20] L. Cheng, Y. Wang, F. Cheng, C. Liu, Z. Zhao, and Y. Wang, "A deep reinforcement learning-based preemptive approach for cost-aware cloud job scheduling," IEEE Transactions on Sustainable Computing, vol. 9, no. 3, pp. 422–432, 2024.

[21] Gowri, S. & Sumathi, A. (2025). Advanced Grasshopper Optimization Scheduling Algorithm (AGOSA) for Multi-Cloud Environments. Journal of Computer Science, 21(10), 2265-2272. https://doi.org/10.3844/jcssp.2025.2265.2272

[22] A. Jayanetti, S. Halgamuge and R. Buyya, "Multi-Agent Deep Reinforcement Learning Framework for Renewable Energy-Aware Workflow Scheduling on Distributed Cloud Data Centers," in IEEE Transactions on Parallel and Distributed Systems, vol. 35, no. 4, pp. 604-615, April 2024, doi: 10.1109/TPDS.2024.3360448.

[23] Gu, Yan & Liu, Zhaoze & Dai, Shuhong & Liu, Cong & Wang, Ying & Wang, Shen & Theodoropoulos, Georgios & Cheng, Long. (2025). Deep Reinforcement Learning for Job Scheduling and Resource Management in Cloud Computing: An Algorithm-Level Review.