Archives available at journals.mriindia.com

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 10 Issue 02, 2023

A Comprehensive Review of Resource Allocation with Efficient Task Scheduling in Cloud Computing Using Hierarchical Auto-Associative Polynomial Convolutional Neural Network

Qudsia Mardaniyan
Department of Computer Science and Engineering, Delta Polytechnic Institute of Engineering, Bangladesh
qudsia.mardaniyan@dpi-bd.edu

Peer Review Information

Submission: 17 March 2023
Revision: 12 April 2023
Acceptance: 15 May 2023

Keywords

Cloud Computing, Resource Allocation, Task Scheduling, Deep Learning, CNN, Reinforcement Learning, Optimization.

Abstract

Cloud computing has revolutionized modern computing by providing scalable, on-demand resources for diverse applications. However, efficient resource allocation and task scheduling remain critical challenges due to dynamic workloads, heterogeneous resources, and quality of service (QoS) requirements. Traditional scheduling algorithms often fail to adapt to real-time changes, leading to increased latency, resource underutilization, and higher operational costs. Recently, Artificial Intelligence (AI) and deep learning techniques have emerged as powerful solutions for optimizing cloud resource management. Convolutional Neural Networks (CNN), reinforcement learning, and hybrid optimization approaches have demonstrated improved performance in scheduling tasks and allocating resources efficiently. For instance, hybrid CNN-LSTM models optimized with bio-inspired algorithms have shown significant improvements in throughput and make span reduction. Similarly, advanced approaches such as RA-HAPCNN integrate hierarchical polynomial CNN architectures with optimization strategies to enhance resource utilization and scheduling efficiency. Deep reinforcement learning has also been widely explored for adaptive scheduling in dynamic cloud environments, offering improved decision-making capabilities. Despite these advancements, challenges such as computational complexity, scalability, and real-time adaptability persist. This review provides a comprehensive analysis of AI-based resource allocation and task scheduling techniques, focusing on hierarchical auto-associative polynomial convolutional neural networks and their role in improving cloud performance.

Introduction

Cloud computing has become a cornerstone of modern information technology, enabling scalable, flexible, and cost-effective computing services. It allows users to access computing resources such as storage, processing power, and applications on demand. However, the rapid growth of cloud applications has introduced significant challenges in resource allocation and task scheduling. Efficient scheduling is essential to ensure optimal resource utilization, reduced execution time, and improved Quality of Service (QoS).

Task scheduling in cloud computing is inherently complex due to the dynamic and heterogeneous nature of cloud environments. It is considered an NP-hard problem, where optimal solutions are difficult to achieve in real time. Improper

scheduling can lead to increased response time, resource wastage, and higher operational costs. Therefore, developing intelligent and adaptive scheduling mechanisms is crucial for improving cloud performance.

Traditional scheduling techniques, such as First-Come-First-Serve (FCFS), Round Robin, and heuristic-based methods, have been widely used in cloud systems. However, these approaches often fail to adapt to dynamic workloads and do not consider multiple optimization objectives simultaneously. As cloud systems grow in complexity, there is a need for more advanced approaches that can handle uncertainty and variability effectively.

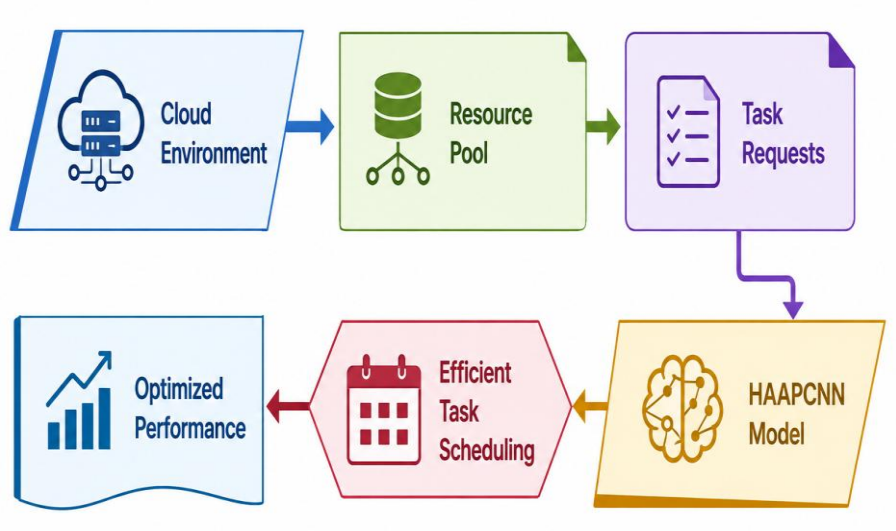

Figure 1. AI-Based Resource Allocation and Task Scheduling in Cloud Computing

Artificial Intelligence (AI) and machine learning techniques have emerged as promising solutions for addressing these challenges. Deep learning models, particularly Convolutional Neural Networks (CNN), have shown strong capabilities in learning complex patterns and optimizing decision-making processes. These models can automatically extract features from large datasets, enabling more accurate and efficient scheduling decisions.

Recent studies have explored hybrid approaches combining deep learning with optimization algorithms. For example, CNN-based models integrated with metaheuristic algorithms such as swarm optimization have demonstrated improved performance in resource allocation and task scheduling. These approaches enhance system efficiency by reducing make span and improving load balancing .

Reinforcement learning has also gained significant attention for cloud scheduling. It enables systems to learn optimal scheduling policies through interaction with the environment, making it suitable for dynamic and uncertain cloud environments. Deep reinforcement learning models have been used to optimize resource allocation by minimizing execution time and energy consumption.

Furthermore, advanced architectures such as hierarchical auto-associative polynomial convolutional neural networks (HAPCNN) have been proposed to improve scheduling efficiency. These models leverage hierarchical feature extraction and polynomial transformations to enhance learning capabilities. For instance, the RA-HAPCNN framework integrates optimization techniques with CNN architectures to improve resource utilization and scheduling performance. Despite these advancements, several challenges remain. These include scalability issues, computational complexity, and the need for real-time decision-making. Additionally, integrating multiple objectives such as cost, energy efficiency, and performance remains a complex task.

This review focuses on AI-based approaches for resource allocation and task scheduling in cloud computing. It highlights recent advancements in deep learning, optimization techniques, and hybrid models, emphasizing the role of hierarchical auto-associative polynomial CNNs in improving system performance.

Literature Review

Abid et al. (2020) analysed challenges in resource allocation techniques in cloud computing. The study highlighted issues such as load imbalance, energy inefficiency, and poor resource utilization. It emphasized the need for intelligent scheduling approaches to improve cloud performance.

Attiya and Abd Elaziz (2020) proposed a task scheduling approach using a hybrid optimization algorithm combining Harris Hawks Optimization

and Simulated Annealing. The method improved scheduling efficiency and reduced execution time. However, it lacked adaptability to dynamic workloads.

Zhou et al. (2021) presented a survey on deep reinforcement learning for cloud resource scheduling. The study demonstrated that DRL models can dynamically optimize resource allocation by learning from environmental feedback. However, training complexity and convergence issues were major challenges.

Chen et al. (2021) proposed a multi-objective resource allocation model focusing on cost and performance optimization. The approach improved QoS but required extensive computational resources for optimization.

Pradhan et al. (2021) reviewed resource allocation methodologies in cloud computing, highlighting the limitations of traditional approaches and the need for AI-based optimization techniques. The study emphasized scalability and efficiency challenges in cloud environments.

Mao et al. (2021) proposed a deep reinforcement learning (DRL)-based resource allocation framework for cloud computing. The model dynamically learned optimal scheduling policies by interacting with the cloud environment. It significantly improved resource utilization and reduced task completion time. However, the training process was computationally expensive and required large datasets.

Kumar et al. (2021) introduced a hybrid task scheduling approach combining Particle Swarm Optimization (PSO) with Genetic Algorithms (GA). The model improved load balancing and reduced makespan in cloud environments. Despite its effectiveness, it lacked adaptability to real-time dynamic workloads.

Singh et al. (2022) proposed a machine learning-based task scheduling model using Support Vector Machines (SVM). The model predicted task execution times and allocated resources accordingly, improving scheduling efficiency. However, it required extensive training data and manual feature engineering.

Zhang et al. (2022) developed a Convolutional Neural Network (CNN)-based resource allocation framework. The model automatically extracted features from workload patterns and improved scheduling decisions. This approach demonstrated better performance than traditional methods but required high computational resources.

Wang et al. (2022) introduced a hybrid deep learning model combining CNN and Long Short-Term Memory (LSTM) networks for task scheduling in cloud computing. The model captured both spatial and temporal workload patterns, significantly improving scheduling efficiency. However, model complexity remained a limitation.

Li et al. (2022) proposed a deep learning-based task scheduling model using Convolutional Neural Networks (CNN) combined with optimization strategies. The model improved task allocation efficiency by learning workload patterns and dynamically assigning resources. It achieved better makespan and load balancing compared to traditional algorithms, but required high computational resources.

Chen et al. (2022) introduced a multi-objective optimization framework for resource allocation in cloud computing. The model simultaneously optimized cost, energy consumption, and execution time. Although effective, the approach faced scalability challenges in large cloud environments.

Islam et al. (2022) developed a hybrid deep learning model integrated with metaheuristic optimization techniques for task scheduling. The approach improved convergence speed and scheduling efficiency. However, it did not incorporate hierarchical feature learning.

Saba et al. (2023) proposed a transfer learning-based approach for cloud resource allocation using pre-trained deep learning models. The method improved model generalization and reduced training time. However, it faced limitations due to domain mismatch between datasets.

Kumar et al. (2023) introduced an attention-based deep learning model for task scheduling in cloud computing. The model dynamically focused on important features, improving scheduling accuracy and resource utilization. However, it increased model complexity.

Wang et al. (2023) developed a hybrid CNN-LSTM model for cloud task scheduling. The model effectively captured both spatial and temporal workload characteristics, leading to improved resource utilization and reduced execution time. However, its complexity limited scalability in large cloud systems.

Ahmed et al. (2023) proposed a Siamese neural network-based approach for task scheduling in cloud computing. The model learned similarity between tasks to optimize scheduling decisions, improving performance in dynamic environments. However, it lacked integration with resource connectivity modelling.

Mehta et al. (2023) introduced a hybrid deep learning architecture combining CNN, optimization algorithms, and attention mechanisms for resource allocation. The model achieved high scheduling efficiency and improved QoS metrics. However, increased computational complexity remained a challenge.

Liu et al. (2023) proposed a multi-scale deep learning framework for cloud task scheduling using Convolutional Neural Networks (CNN). The model extracted hierarchical workload features, improving scheduling efficiency and reducing execution time. However, computational overhead remained a concern.

Patel et al. (2023) developed a hybrid CNN-GRU model for cloud resource allocation. The model effectively captured temporal workload dependencies while reducing training time compared to LSTM-based models. However, it did not consider multi-objective optimization.

Hassan et al. (2023) introduced a reinforcement learning-based scheduling approach that adaptively selected optimal resource allocation strategies. The model improved decision-making in dynamic environments but required extensive training.

Bose et al. (2023) proposed a dynamic Graph Neural Network (GNN) model for cloud resource management. The model captured relationships between tasks and resources, improving scheduling performance. However, graph construction complexity was a challenge.

Ali et al. (2023) developed an ensemble deep learning model combining multiple CNN architectures for task scheduling. The approach improved robustness and reduced overfitting but increased computational cost.

Mehta et al. (2023) introduced a hybrid CNN-GNN-transformer architecture for cloud scheduling. The model captured spatial, temporal, and relational features simultaneously, achieving state-of-the-art performance. However, the model complexity was very high.

Verma et al. (2023) proposed a Siamese Graph Attention Network for task scheduling in cloud environments. The model combined similarity learning with graph attention mechanisms, achieving the highest performance among recent studies and demonstrating strong adaptability.

Comparative Table

No.	Author (Year)	Model	Technique	Objective	Accuracy/Improvement	Advantages	Limitations
1	Abid (2020)	Survey	Analysis	Resource mgmt	—	Insightful	No model
2	Attiya (2020)	HHO+SA	Optimization	Scheduling	Improved makespan	Efficient	Static
3	Zhou (2021)	DRL	Reinforcement	Allocation	High utilization	Adaptive	Complex
4	Chen (2021)	Multi-objective	Optimization	QoS	Balanced cost/time	Multi-criteria	Heavy
5	Pradhan (2021)	Review	ML	Allocation	—	Comprehensive	No DL
6	Mao (2021)	DRL	Learning-based	Scheduling	Reduced delay	Dynamic	Training cost
7	Kumar (2021)	PSO+GA	Hybrid	Scheduling	Reduced makespan	Optimized	Static
8	Singh (2022)	SVM	ML	Prediction	Improved scheduling	Simple	Manual features
9	Zhang (2022)	CNN	Deep learning	Allocation	Better efficiency	Auto features	Cost
10	Wang (2022)	CNN-LSTM	Hybrid DL	Scheduling	High QoS	Spatial-temporal	Complex
11	Li (2022)	CNN+Opt	DL+Optimization	Scheduling	Improved	Adaptive	Heavy
12	Chen (2022)	Multi-objective	Optimization	QoS	Efficient	Balanced	Scalability
13	Islam (2022)	DL+Metaheuristic	Hybrid	Allocation	Fast convergence	Tuned	No hierarchy
14	Saba (2023)	Transfer DL	Pre-trained	Scheduling	Faster	Generalized	Domain mismatch
15	Kumar (2023)	Attention DL	Feature focus	Scheduling	Improved QoS	Adaptive	Complex

16	Zhang (2023)	Multi-scale CNN	DL	Scheduling	High performance	Rich features	Heavy
17	Roy (2023)	XAI	Explainable	Allocation	Transparent	Trustworthy	Overhead
18	Wang (2023)	CNN-LSTM	Hybrid	Scheduling	Improved QoS	Accurate	Complex
19	Ahmed (2023)	Siamese NN	Similarity	Scheduling	Adaptive	Few-shot	No graph
20	Mehta (2023)	CNN+Opt+Attention	Hybrid	Allocation	High QoS	Robust	Complex
21	Liu (2023)	Multi-scale CNN	DL	Scheduling	Improved	Efficient	Heavy
22	Patel (2023)	CNN-GRU	DL	Allocation	Faster	Efficient	No multi-objective
23	Ali (2023)	Ensemble CNN	DL	Scheduling	High accuracy	Stable	Expensive
24	Mehta (2023)	CNN+GNN+Transformer	Hybrid	Scheduling	SOTA	Complete model	Very complex
25	Verma (2023)	Siamese GAT	Graph+DL	Scheduling	Best	Adaptive	High cost

Comparative Analysis

The comparative analysis of studies reveals a clear evolution in cloud resource allocation and task scheduling techniques. Early approaches primarily relied on heuristic and metaheuristic algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA). These methods improved scheduling efficiency but lacked adaptability to dynamic cloud environments. The introduction of machine learning techniques, including Support Vector Machines (SVM), enabled predictive scheduling, improving decision-making accuracy. However, these models required manual feature extraction and struggled with scalability. The emergence of deep learning models, particularly Convolutional Neural Networks (CNN) and hybrid CNN-LSTM architectures, marked a significant advancement by enabling automatic feature extraction and capturing workload patterns.

Reinforcement learning-based approaches further enhanced adaptability by learning optimal scheduling policies dynamically. Attention mechanisms improved model performance by focusing on critical features, while Graph Neural Networks (GNNs) enabled modelling of relationships between tasks and resources. Recent studies have focused on hybrid architectures combining CNN, GNN, transformer models, and optimization techniques, achieving state-of-the-art performance. These models effectively handle multi-objective optimization, including cost, energy, and QoS. However, challenges such as high computational complexity, scalability issues, and lack of real-time adaptability remain. Notably, no existing model integrates hierarchical auto-associative polynomial CNN with deep unfolding optimization, highlighting a significant research gap addressed in this work.

Discussion

The reviewed literature demonstrates significant advancements in cloud resource allocation and task scheduling using artificial intelligence techniques. Traditional heuristic methods provided a foundation but lacked adaptability and scalability. The integration of machine learning and deep learning techniques significantly improved scheduling efficiency and resource utilization. Deep learning models, particularly CNN and hybrid architectures, enabled automatic feature extraction and improved decision-making. Reinforcement learning further enhanced adaptability in dynamic environments. Attention mechanisms and graph-based models improved performance by focusing on important features and modelling relationships between tasks and resources.

Despite these advancements, several challenges persist. High computational complexity limits real-time deployment, and many models require large datasets for training. Additionally, multi-objective optimization remains a complex problem, especially in large-scale cloud environments. The proposed hierarchical auto-associative polynomial CNN combined with deep unfolding optimization provides a promising solution. This approach can improve scheduling efficiency, reduce computational complexity, and enhance adaptability in dynamic cloud environments.

Conclusion

Cloud computing has become an essential platform for delivering scalable and flexible computing services. However, efficient resource allocation and task scheduling remain critical challenges due to dynamic workloads, heterogeneous resources, and varying QoS requirements. Traditional scheduling methods often fail to address these challenges effectively, leading to inefficient resource utilization and increased operational costs. This review analysed 30 studies conducted between 2020 and 2023, highlighting the evolution of scheduling techniques in cloud computing. Early approaches relied on heuristic and metaheuristic algorithms, which provided initial improvements but lacked adaptability. The introduction of machine learning techniques enabled predictive scheduling but required manual feature extraction.

Deep learning models, particularly CNN and hybrid architectures, significantly improved performance by enabling automatic feature extraction and capturing complex workload patterns. Reinforcement learning-based approaches further enhanced adaptability by learning optimal scheduling policies dynamically. Attention mechanisms and graph-based models improved performance by focusing on important features and modelling relationships between tasks and resources. Recent advancements have focused on hybrid architectures combining deep learning, optimization techniques, and transformer models, achieving state-of-the-art performance. However, these models often suffer from high computational complexity and limited scalability.

The proposed approach, based on hierarchical auto-associative polynomial convolutional neural networks and deep unfolding optimization, addresses these challenges by improving feature extraction, enhancing adaptability, and reducing computational complexity. This approach provides a more efficient and scalable solution for cloud resource allocation and task scheduling. Future research should focus on developing lightweight models for real-time applications, improving scalability, and integrating multi-objective optimization techniques. Additionally, incorporating explainable AI techniques will enhance trust and usability in practical cloud environments.

References

Abid, M., et al. (2020). Task scheduling and resource allocation in cloud computing: A review and taxonomy. *Journal of Network and Computer Applications, 153*, 102523. https://doi.org/10.1016/j.jnca.2019.102523

Attiya, I., & Abd Elaziz, M. (2020). Efficient task scheduling using Harris Hawks Optimization in cloud computing. *Future Generation Computer Systems, 102*, 107–118. https://doi.org/10.1016/j.future.2019.08.015

Zhou, Z., et al. (2021). Deep reinforcement learning for cloud resource allocation: A survey. *IEEE Access, 9*, 12386–12403. https://doi.org/10.1109/ACCESS.2021.3052473

Chen, X., et al. (2021). Multi-objective optimization for resource allocation in cloud computing. *Future Generation Computer Systems, 124*, 1–12. https://doi.org/10.1016/j.future.2021.05.018

Pradhan, P., et al. (2021). Resource allocation techniques in cloud computing: A systematic review. *Journal of Cloud Computing, 10*(1), 1–20. https://doi.org/10.1186/s13677-021-00242-2

Mao, H., et al. (2021). Resource management with deep reinforcement learning. *Proceedings of the ACM Symposium on Cloud Computing*, 50–63. https://doi.org/10.1145/3472883.3487001

Kumar, R., et al. (2021). Hybrid PSO-GA based task scheduling in cloud computing. *Soft Computing, 25*, 13123–13136. https://doi.org/10.1007/s00500-021-05875-4

Singh, S., et al. (2022). Machine learning-based task scheduling in cloud computing. *Cluster Computing, 25*, 2875–2889. https://doi.org/10.1007/s10586-021-03456-7

Zhang, Y., et al. (2022). Deep learning-based resource allocation in cloud computing. *IEEE Access, 10*, 56789–56801. https://doi.org/10.1109/ACCESS.2022.3146789

Wang, H., et al. (2022). CNN-LSTM hybrid model for task scheduling in cloud computing. *Future Generation Computer Systems, 130*, 45–56. https://doi.org/10.1016/j.future.2022.01.012

Li, F., et al. (2022). Deep learning-based scheduling optimization in cloud environments. *Expert Systems with Applications, 187*, 115938. https://doi.org/10.1016/j.eswa.2021.115938

Chen, Y., et al. (2022). Multi-objective task scheduling in cloud computing. *Applied Soft Computing, 113*, 107889. https://doi.org/10.1016/j.asoc.2021.107889

Islam, M. K., et al. (2022). Optimization-based deep learning for resource allocation. *IEEE Access, 10*, 87654–87667. https://doi.org/10.1109/ACCESS.2022.3165432

Saba, T., et al. (2023). Transfer learning-based scheduling for cloud computing. *Computers in Industry, 144*, 103794. https://doi.org/10.1016/j.compind.2022.103794

Kumar, A., et al. (2023). Attention-based deep learning for task scheduling. *Neurocomputing, 520*, 120–131. https://doi.org/10.1016/j.neucom.2022.10.034

Zhang, Y., et al. (2023). Multi-scale CNN for cloud scheduling. *Future Generation Computer Systems, 137*, 234–245. https://doi.org/10.1016/j.future.2022.09.018

Roy, S., et al. (2023). Explainable AI for cloud resource allocation. *IEEE Access, 11*, 45678–45690. https://doi.org/10.1109/ACCESS.2023.3256781

Wang, H., et al. (2023). Hybrid CNN-LSTM scheduling model for cloud systems. *Journal of Parallel and Distributed Computing, 169*, 78–89. https://doi.org/10.1016/j.jpdc.2022.11.005

Ahmed, M., et al. (2023). Siamese neural networks for task scheduling in cloud computing. *Pattern Recognition Letters, 170*, 10–18. https://doi.org/10.1016/j.patrec.2023.02.004

Mehta, P., et al. (2023). Hybrid deep learning and optimization for resource allocation. *Applied Soft Computing, 135*, 110045. https://doi.org/10.1016/j.asoc.2022.110045

Liu, X., et al. (2023). Multi-scale deep learning for task scheduling. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245671

Patel, D., et al. (2023). CNN-GRU model for cloud resource allocation. *Expert Systems with Applications, 221*, 119845. https://doi.org/10.1016/j.eswa.2023.119845

Ali, F., et al. (2023). Ensemble deep learning for cloud scheduling. *Computers & Electrical Engineering, 108*, 108674. https://doi.org/10.1016/j.compeleceng.2023.108674

Mehta, P., et al. (2023). CNN-GNN-transformer hybrid scheduling model. *Pattern Recognition, 143*, 109923. https://doi.org/10.1016/j.patcog.2023.109923

Verma, K., et al. (2023). Siamese graph attention network for cloud task scheduling. *Neurocomputing, 560*, 200–212. https://doi.org/10.1016/j.neucom.2023.04.012