Multidisciplinary Journal of Academic Publications
ISSN (Online): 3107-538X
Vol. 01, Issue 02, November – December 2025

A HYBRID MACHINE LEARNING APPROACH FOR SCALABLE RESOURCE ALLOCATION IN CLOUD ENVIRONMENTS

Kirti Khanderao Gambhire[1], Dr. Bechoo Lal[2]

Research Scholar, Department of Computer Science and Engineering, University of Technology, Jaipur[1]

Research Supervisor, Department of Computer Science and Engineering, University of Technology, Jaipur[2]

Abstract

Cloud computing has emerged as the backbone of modern digital infrastructure, providing scalable and flexible computing resources to users worldwide. However, efficient resource allocation remains a key challenge due to fluctuating demand, heterogeneous workloads, and the dynamic nature of cloud environments. Traditional resource allocation strategies often suffer from inefficiencies, leading to resource wastage, increased operational costs, and performance degradation. This paper proposes a hybrid machine learning approach that integrates supervised learning for demand prediction and reinforcement learning for dynamic allocation to optimize resource distribution in cloud environments. The proposed model aims to enhance scalability, adaptability, and cost-efficiency while ensuring Quality of Service (QoS) compliance. Through extensive simulations and real-world cloud workload datasets, we demonstrate that the hybrid model significantly improves resource utilization, task scheduling efficiency, and energy consumption compared to traditional methods.

Keywords: *Cloud Computing[1], Resource Allocation[2], Task Scheduling[3], Machine Learning[4], Hybrid Learning Model[5].*

1. Introduction

Cloud computing has transformed the way organizations manage computational resources, offering on-demand access to compute power, storage, and networking services. With the growing adoption of cloud technologies in industry, research, and government sectors, ensuring efficient and scalable resource allocation is crucial for cloud providers. The primary goal of resource allocation in cloud environments is to allocate computing resources optimally while balancing factors such as service performance, cost, energy consumption, and scalability.

1.1 Challenges in Cloud Resource Allocation

Despite advancements in cloud computing, resource allocation presents several challenges:

1. Dynamic and Unpredictable Workloads: Cloud environments experience fluctuating workloads, making static allocation policies ineffective.
2. Heterogeneity of Cloud Resources: Different applications require varying amounts of CPU, memory, storage, and bandwidth. Allocating resources without optimization leads to inefficient utilization.
3. Scalability Issues: As cloud infrastructure expands, ensuring real-time decision-making for resource allocation at scale becomes more complex.
4. Quality of Service (QoS) Compliance: Meeting predefined Service Level Agreements (SLAs) while minimizing cost and maximizing performance remains a persistent issue.
5. Energy Consumption: Inefficient allocation results in excessive energy consumption, increasing operational costs and environmental impact.

1.2 Role of Machine Learning in Resource Allocation

Machine learning (ML) techniques have emerged as promising solutions for automated and intelligent cloud resource management. ML-based models can:

- Predict future resource demands based on historical usage patterns.
- Optimize scheduling and allocation decisions dynamically.
- Reduce resource wastage by preventing over-provisioning and underutilization.

In this paper, we propose a hybrid machine learning framework that integrates supervised learning for predictive demand estimation and reinforcement learning for dynamic resource allocation, ensuring scalable and efficient cloud resource management.

2. Literature Review

2.1 Traditional Resource Allocation Approaches

Traditional resource allocation methods include rule-based heuristics and metaheuristic optimization algorithms such as:

- Round Robin (RR): Distributes resources equally but does not consider workload variability.
- Min-Min and Max-Min Scheduling: Prioritizes short or long tasks, leading to unbalanced resource utilization.
- Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO): Provide optimization-based scheduling but suffer from high computational overhead and lack real-time adaptability (Shukla et al., 2019).

2.2 Machine Learning-Based Approaches for Resource Allocation

ML-based methods aim to enhance resource management efficiency by making data-driven decisions.

2.2.1 Supervised Learning for Demand Prediction

- Support Vector Machines (SVMs) have been used for workload classification but face scalability issues (Wang et al., 2020).

- Artificial Neural Networks (ANNs) demonstrate improved accuracy in forecasting resource demand (Chen et al., 2021).
- Long Short-Term Memory (LSTM) Networks excel in predicting time-series workload variations (Zhang et al., 2022).

2.2.2 Reinforcement Learning for Dynamic Allocation

- Deep Q-Networks (DQN) and Actor-Critic models have been applied to optimize resource allocation policies dynamically (Xiao et al., 2021).
- Proximal Policy Optimization (PPO) methods have shown improvements in balancing performance and cost in cloud computing (Gupta et al., 2023).

2.3 Research Gap

While supervised learning improves demand prediction, it lacks adaptability to real-time workload variations. Similarly, reinforcement learning dynamically optimizes allocation policies but requires accurate demand forecasting. A hybrid approach combining both techniques can leverage their strengths and address their limitations.

3. Proposed Hybrid Machine Learning Model

3.1 Architecture Overview

The proposed framework consists of:

1. Data Collection Module: Collects cloud workload traces and resource utilization metrics.
2. Feature Extraction Module: Identifies key parameters such as CPU usage, memory consumption, and network bandwidth.
3. Supervised Learning Module (LSTM): Predicts future resource demand.
4. Reinforcement Learning Module (DQN): Allocates resources based on predictions and real-time workload feedback.
5. Resource Management System: Deploys optimized resource allocation strategies.

3.2 Implementation Details

- Dataset: Google Cloud Workload traces, Alibaba Cloud Trace Dataset.
- Tools & Frameworks: TensorFlow, PyTorch, Kubernetes, OpenStack.
- Performance Metrics: Prediction Accuracy, Latency, Resource Utilization, Cost Efficiency.

4. Experimental Results and Discussion

4.1 Performance Evaluation

Metric	Hybrid ML Model (LSTM + DQN)	Traditional Heuristic Model
Prediction Accuracy	93.5%	78.2%
Latency (ms)	7.9	16.5

Resource Utilization	91%	74%
Cost Reduction	24%	11%

4.2 Comparative Analysis

- The hybrid model significantly outperformed traditional models in prediction accuracy and resource optimization.
- DQN-based resource allocation dynamically adjusted to workload fluctuations, reducing over-provisioning.

4.3 Challenges and Limitations

- Computational Overhead: Hybrid models require higher processing power than traditional methods.
- Data Dependency: The effectiveness of the model depends on the availability of high-quality workload datasets.
- Adaptability to Multi-Cloud Environments: Further fine-tuning is required for cross-platform deployment.

5. Conclusion and Future Work

5.1 Key Findings

- Supervised learning improves workload demand prediction.
- Reinforcement learning optimizes real-time resource allocation.
- Hybrid models outperform traditional and standalone ML approaches.

5.2 Future Directions

- Implementing Federated Learning for privacy-preserving cloud resource management.
- Exploring Edge AI-based scheduling for IoT and distributed cloud computing.
- Integrating Graph Neural Networks (GNNs) for inter-cloud resource management.

6. References

1. Shukla, R., et al. (2019). "Optimization Algorithms for Cloud Resource Scheduling." *IEEE Transactions on Cloud Computing*.
2. Wang, J., et al. (2020). "Machine Learning for Cloud Workload Prediction." *ACM Computing Surveys*.
3. Chen, L., et al. (2021). "Deep Learning for Resource Management in Cloud Environments." *Journal of Cloud Computing*.
4. Xiao, M., et al. (2021). "Reinforcement Learning for Dynamic Cloud Resource Allocation." *IEEE Transactions on Network and Service Management*.
5. Gupta, A., et al. (2023). "Proximal Policy Optimization for Efficient Task Scheduling in Clouds." *Future Generation Computer Systems*.