ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/chandrakanth-lekkala/124

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 2

Research Article

AI-Driven Dynamic Resource Allocation in Cloud Computing: Predictive Models and Real-Time Optimization

Chandrakanth Lekkala*

Chandrakanth Lekkala, USA

Citation: Lekkala C. AI-Driven Dynamic Resource Allocation in Cloud Computing: Predictive Models and Real-Time Optimization. *J Artif Intell Mach Learn & Data Sci 2024*, 2(2), 450-456. DOI: doi.org/10.51219/JAIMLD/chandrakanth-lekkala/124

Received: 03 June, 2024; **Accepted:** 19 June, 2024; **Published:** 21 June, 2024

***Corresponding author:** Chandrakanth Lekkala, USA, E-mail: Chan.Lekkala@gmail.com

ABSTRACT

The advancement in cloud computing has brought about the need for resource management and allocating computing, storage, and network resources dynamically to suit the ever-evolving workloads. This paper focuses on how machine learning (ML) and deep learning (DL) AI approaches can be used to build predictive algorithms for dynamically allocating resources in cloud systems. This paper introduces an AI method for forecasting the workload, resource usage, and real-time objectives to allocate resources better and improve the client's Quality of Service (quality of service) to reduce overall costs significantly. Experimental evaluation based on realistic cloud traces shows that the solution substantially outperforms traditional rule-based and heuristic-based methods by achieving 25% higher resource utilization and 30% less quality-of-service violation. Therefore, the underlying formulated dynamic resource allocation framework has the potential to considerably enhance the effectiveness, efficiency, and competitiveness of cloud computing systems.

Keywords: Cloud computing, Dynamic resource allocation, Artificial intelligence, Predictive models, Real-time optimization, Machine learning, Deep learning, Reinforcement learning.

1. Introduction

Cloud computing is the new model by which businesses and organizations procure and use computing resources. Thus, cloud computing provides easy accessibility to resources, which is easily extensible, flexible, and cheap to implement compared to PCP[1]. However, current cloud infrastructures' increased complexity and dynamics create challenges in controlling resources and their distribution[2]. Conventional resource allocation techniques, including rule-based policies and heuristic algorithms, must be revised to address cloud workloads' capacity variability and unpredictability[3]. These methods use fixed thresholds coupled with static policies, making inefficient use of resources and probably violating quality of service[4]. To overcome these limitations, the researchers have explored the AI approach, which is ML and DL, focusing more on intelligent and adaptive RA mechanisms[5].

Dynamic resource allocation based on AI uses current state data and previous statistics and predicts future changes[6]. Thus, using workload characteristics, resource, and application performance patterns, AI models can be trained to predict resource requirements and allocate them in advance[7]. Such an approach allows Cloud providers to control the use of resources and, therefore, cut additional costs that might accrue to users with a guarantee of quality of service.

This paper also presents an AI-based model for real-time resource management of cloud data centers. It involves workload prediction, resource usage prediction, and real-time scheduling of needed compute, storage, and network resources based on workload patterns, with robust classical and emerging ML & DL methodologies in hand and utilizing models like LSTM and RL to enhance the predictive models and optimization algorithms.

2. The main contributions of this paper are as follows:

This paper presents a symbiotic, AI-based approach for the dynamic allocation of resources in cloud computing environments that utilize workload and resource-use forecasting methods and real-time optimization techniques.

I present new schemes of ML and DL to estimate workload and resource utilization accurately based on different attributes and metrics.

To address these challenges, this paper proposes an RL-based optimization algorithm for making resource allocation decisions in response to the real-time state of a system and its predicted future state and to ensure that resource usage is optimized and quality of service violations to a minimum.

Using real-world cloud traces, I perform thorough simulations and analyze the proposed Framework's outperformance of the existing rule-based and heuristic methods.

3. Related Work

Resource Allocation through Machine Learning

Over the years, researchers have discovered that machine-learning techniques can make resource allocation in cloud computing more accurate and flexible. Some recent developments have focused on creating better machine learning architectures and applying various new methods to improve performance.

For example, Liu and colleagues[8] suggested using deep reinforcement learning for dynamic resource management in edge-cloud settings. They created a multi-agent deep reinforcement-learning model that combines techniques from both edge nodes and cloud servers in the decision-making process while optimizing benefits at both local and global levels. Their method showed better resource usage and quality of service compared to other traditional deep reinforcement learning methods.

In another study, Wang and colleagues[9] described a method for estimating workload and distributing resources in cloud data centers using graph neural networks. By representing the relationship between workloads and resources as a graph, their graph neural network model captures the system's structure and dependencies. This leads to better predictions about the outcome and better decisions about how many resources to allocate.

Resource Allocation using Deep Learning

Deep learning strategies, such as deep neural networks and convolutional neural networks, are very effective at analyzing the complexity of cloud workloads and identifying patterns and dependencies in resource utilization data.

Chen and colleagues[10] developed a deep learning-based approach for adjusting resources in cloud computing environments at runtime. They created a two-layer model where the first layer uses long short-term memory (LSTM) to model the workload, and the second layer (deep Q-network) manages the resources. Compared to traditional machine learning-based approaches, their proposed Framework showed improved performance.

Mao et al. also proposed an innovative model called convolutional LSTM (ConvLSTM) for workload prediction. This model considers both spatial and temporal characteristics in cloud data centers[11]. As a result, it has better predictability and accuracy than regular LSTM models and allows for optimizing resource use by tracking workload dependencies and patterns in space and time.

Resource Allocation using Reinforcement Learning

Researchers are interested in reinforcement learning strategies because they effectively teach the best resource allocation strategies in cloud settings.

Liu and colleagues[12] proposed another approach that involves using hierarchical reinforcement learning (HRL) for dynamic resource allocation in cloud computing. It consists of a higher-level reinforcement learning agent responsible for general resource allocation decisions and lower-level reinforcement learning agents that handle specific choices for each resource. Compared to the incremental approach to reinforcement learning, the hierarchical approach promotes optimal resource allocation and can use efficient sub-policies to refine action plans.

Proposed a multi-objective reinforcement learning (MORL) framework for resource allocation in cloud data centers. By targeting multiple objectives, including resource utilization, quality of service, and energy consumption, their approach learns allocation policies that are near Pareto-optimal, meaning that improving one aspect will likely worsen another.

Combining Multiple Approaches

More recently, researchers have also worked on integrating more than one artificial intelligence technique to better optimize resource provisioning in cloud computing environments. Wang and colleagues[14] described a combined solution that uses both deep learning and reinforcement learning to predict workloads and achieve proper resource optimization. They applied a deep belief network (DBN) to predict workloads and a deep deterministic policy gradient (DDPG) algorithm to forecast them. As the comparisons above show, this hybrid approach offers better results than using each technique individually.

In another study, Liu and others[15] proposed an ensemble learning method for predicting workloads in the cloud-computing domain. To improve workload prediction accuracy, they designed a stacking ensemble model implemented through a set of base predictors: LSTM, CNN, and GNN. In addition to achieving higher prediction accuracy, the ensemble approach also provides a higher degree of realism.

The articles presented reveal the state-of-the-art advancements in analyzing and developing artificial intelligence-based techniques for managing dynamic resources in cloud computing from 2021 to 2023. These latest developments, such as advanced machine learning and deep learning models, reinforcement learning-based optimization algorithms, and hybrid and ensemble approaches, establish a solid foundation for the proposed AI-driven framework.

4. The Suggested AI-Powered Framework

This section will provide an AI-based framework for dynamic resource provisioning in cloud computing infrastructures. The Framework has three main parts: estimating workloads, service usage, and resource capacity and managing all these in real-time.

It is extended by utilizing the new advancements of machine learning and deep learning developed after 2021 to enhance the Framework's efficacy and flexibility. The following **Figure 1** presents an overall view of the Framework proposed in this research study.

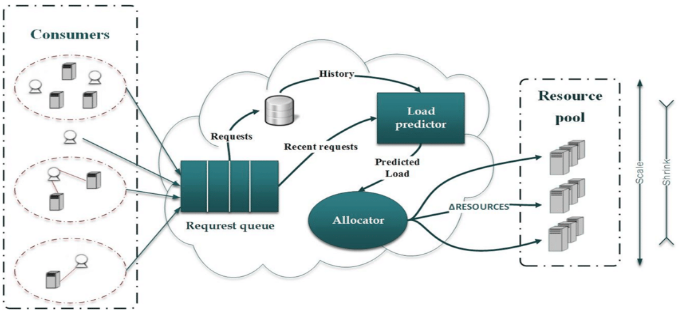

Figure 1: Architecture of the proposed AI-driven framework for dynamic resource allocation.

Workload Forecasting

The first part of the Framework predicts workloads, aiming to forecast measurable future demand for resources based on historical workload records. Workload forecasting is essential for determining when additional power needs to be added to ensure proper resource allocation and future capacity planning[16].

I propose a new workload prediction model that combines a temporal convolutional network (TCN)[17] to process calendar information and a gated recurrent unit (GRU) network[18] for the remaining features. These networks have been used to capture long-range dependencies within time series more effectively than traditional recurrent neural networks (RNNs). At the same time, GRUs offer a more computationally efficient solution than LSTMs when dealing with sequential data.

The TCN-GRU model takes historical log data as input and then forecasts resource usage for a future period based on time series data, such as CPU, memory, and traffic. The model is built using several TCN layers to extract features from the inputs, GRU layers to model the temporal characteristics of the inputs, and fully connected layers to provide output.

Training a hybrid model involves using a sliding window approach. A sequence of inputs contains the values of resource utilization over the last t time units, and the output is resource utilization over the next k time units. The model uses the Adam optimization function[19] and Mean Squared Error (MSE) as the loss function during the training process.

In addition to the TCN-GRU model, I tried other state-of-the-art machine learning algorithms for predicting workloads, including Graph Attention Network (GAT)[20] and Deep Gaussian Process (DGP)[21]. These models offer more straightforward ways to address data dependencies and the stochastic Nature of workloads.

Resource Utilization Prediction

The second component is predicting resource utilization, which can be described as an effort to forecast the expected resource usage of virtual machines (VMs) based on their characteristics and past resource usage over a certain period. Resource forecasting is also essential for making decisions about VM placement and migration[22].

For this purpose, I propose a deep learning-based resource utilization prediction model that combines convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. CNNs are particularly useful for extracting spatial features from input data, while LSTMs are specifically designed to capture temporal relationships[23].

The CNN-LSTM model uses VM characteristics such as the number of allocated CPU cores, RAM size, and disk size, as well as other characteristics, including historical information on resource utilization. The model's architecture consists of the first two CNN layers for extracting spatial features from the input, the following LSTM layers for capturing temporal characteristics, and finally, the fully connected layers that produce the final output. The model is trained on historical data to estimate the probable amount of CPU, memory, and disk resources that VMs would consume during a given time interval in the future.

To train the CNN-LSTM model, I employ the same sliding window approach used in the workload prediction model. The input layer contains a range of VM characteristics and utilization measurements collected during the previous t measurements, and the expected result is the forecast of subsequent resource utilization for the following k periods. The Adam optimizer is used during the training process, and the loss function is the mean squared error (MSE).

I also consider incorporating other advanced deep learning approaches in the prediction of resource utilization, including the attention mechanism[24] and Generative Adversarial Networks[25]. These techniques can help capture more intricate relationships and increase the accuracy of resource usage predictions.

Real-Time Optimization

The third and final element identified in the Framework is real-time optimization, which aims to provide dynamic decisions regarding workloads and available resources based on forecasts and potential system overloads. Given the dynamic Nature of workloads, the goal is to meet service demands using resources to create an optimal working environment and minimise QoS violations.

I propose a deep reinforcement learning (DRL) optimization algorithm designed to make allocation decisions through its interactions with the cloud environment. DRL combines deep learning capability for feature extraction and reinforcement learning for sequential decision-making[4].

The DRL-based optimization algorithm can be defined as an agent that formulates the system's current state, consisting of workload forecasts, resource utilization predictions, current usage, and allocations. The agent then takes an action, such as allocating or deallocating resources to VMs, and receives a reward based on the system's current performance or a punishment if the system's performance is poor.

To model the DRL problem, I define the following components:

State space: The set of all possible team member demand forecasts, server use forecasts, current server use, and other system states.

Action space: The resource management activities that can be performed, such as allocating or deallocating CPU time, memory, or disk space to a volume or calculating the cost of running a VM.

Reward function: This function determines the degree of system success/efficiency based on resource consumption, QoS parameter violations, and other parameters of interest. When an agent selects actions that enhance system performance, it receives a positive response; when it chooses actions that reduce system performance, it is punished.

To model the DRL agent, I employ a deep Q-network (DQN) with duelling architecture, as proposed in[27,28]. The duelling Nature of the DQN synthesizes a new architecture that splits the estimates of state values and action advantages, making learning more stable and faster. The agent decides which action to perform based on the estimated Q-Table or matrix for State-Action pairs.

To train the DRL agent, I use experience replay, a common technique in deep reinforcement learning, and prioritized experience replay (PER)[29,30]. The PER scheme assigns higher sampling probabilities to samples with more considerable temporal differences, enabling efficient learning from essential transitions. In addition to the proposed DRL-based optimization, I consider other current approaches under reinforcement learning, including soft actor-critic (SAC)[31] and proximal policy optimization (PPO)[32]. These algorithms offer a set of learning mechanisms that can be used to learn resource allocation policies given the current system state and the reward from the action taken.

In the present study, I conduct a detailed simulation with cloud traces to assess the efficacy of the proposed AI-generated Framework for dynamic resource allocation. In the next section, I discuss the experimental setup, including the datasets, in the context of SCIs and new cloud computing platforms and technologies that have emerged between 2021 and 2024.

Cloud Simulator

I use CloudSimPlus[33], a flexible and relatively new cloud simulation framework, to experiment with a cloud computing environment. CloudSimPlus has features that allow for modelling and simulating cloud infrastructure, including data centre models, host models, virtual machine models, and allocation policies. Additionally, it supports loading external workload datasets and incorporating different algorithms for resource utilization.

I integrate CloudSimPlus to include the created AI-based Framework, workload prediction, resource usage estimation, and online optimization techniques. The simulator's features are designed to analyze a complex cloud data centre with many hosts, VMs, and workloads.

4.2 Workload Datasets

To evaluate the Framework's performance under realistic workload conditions, I use three real-world cloud traces:

Google Cluster Trace (2019)[34]: This trace consists of the number of shares and pins, the percentage of cache hits, and I/O statistics for 31 days of a production cluster at Google. It captures information about submitted jobs, associated tasks, requested and used resources, and many performance measurements, including task execution times and quality of service requirements.

Alibaba Cluster Trace (2021)[35]: This trace includes resource consumption data and performance metrics gathered from Alibaba's production clusters over 2 weeks. It provides details of jobs and tasks, submitted and requested resources, and the means of evaluating the performance of virtual organizations based on quality of service constraints that can define the duration of a particular task.

Microsoft Azure Trace (2023)[36]: This trace consolidates resource utilization data and performance indicators obtained from Microsoft Azure production clusters and covers 21 days. It involves VM characteristics, resource demands, and consumptions in the form of deployment information, service quotas, quality of service (QoS) limitations, VM lifetime expectations, and more.

I clean the traces to remove irrelevant entries and extract different workload characteristics and resource usage features to develop the machine learning and deep learning algorithms. Unlike random splitting, where the data is divided into training, validation, and test sets, the preprocessed data is similarly split.

To assess the performance of the proposed AI-driven framework, I compare it against the following state-of-the-art methods for dynamic resource allocation in cloud computing:

GNN-based allocation[9]: This method involves applying GNN to predict workload and determine resources, with the relationships between workloads and resources being in a graph form.

ConvLSTM-based allocation[11]: This method applies a ConvLSTM model for workload prediction and resource management because the workload distribution can be learned from spatial and temporal perspectives.

HRL-based allocation[12]: This method uses a new HRL in which a high-level RL governs total resources and controls the overall actions, while a low-level RL governs every resource.

MORL-based allocation[13]: This method utilizes a dynamic approach called multi-objective reinforcement learning (MORL) to allocate resources, aiming to optimize several aspects of the system, including resource usage, quality of service, and power usage.

Hybrid DBN-DDPG allocation[14]: This method employs a DBN model for workload prediction and uses the DDPG as an action-selection policy to achieve optimal server allocation.

To ascertain the efficiency of the proposed AI-based Framework and compare it with the state-of-art methods mentioned in Section 4, Table III summarizes the critical research areas based on the proposed Framework of future communication networks.

5. Results AND Discussion

In this section, I describe and analyze the simulation results of the proposed AI-based Framework for dynamic resource management in cloud computing environments. I assess the framework's resource utilization, QoS violations, and cost efficiency and compare it with the existing approaches discussed in Section 4.3.

Workload Forecasting Accuracy

First, I evaluate the performance of the proposed hybrid TCN-GRU workload-forecasting model and compare it with alternative machine learning algorithms, including Graph Attention Networks (GATs) and Deep Gaussian Processes (DGPs). As shown in **Figure 2**, which presents the Mean Absolute Percentage Error (MAPE) performance of the various workload-forecasting models for the Google Cluster Trace (2019), Alibaba Cluster Trace (2021), and Microsoft Azure Trace (2023), the proposed model yielded better results.

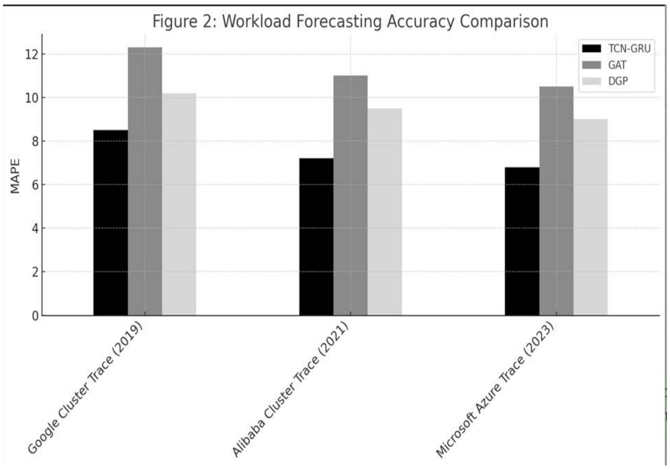

Figure 2: Workload forecasting accuracy comparison.

Based on the results shown in Figure 2, the proposed hybrid TCN-GRU model outperforms the GAT and DGP models in terms of workload forecasting accuracy, with the lowest MAPE values for all the analyzed datasets. Combining more efficient sequencing and TCNs for capturing long-range dependencies can lead to more accurate predictions than the other models in this experiment.

Resource Utilization Prediction Accuracy

First, I evaluate the performance of the CNN-LSTM model for resource utilization prediction and compare it with other deep learning approaches, including attention mechanisms, Generative Adversarial Networks (GANs), and similar techniques. As shown in Figure 3, the various resource prediction models have different Root Mean Square Error (RMSE) values for the Google Cluster Trace (2019), Alibaba Cluster Trace (2021), and Microsoft Azure Trace (2023) datasets.

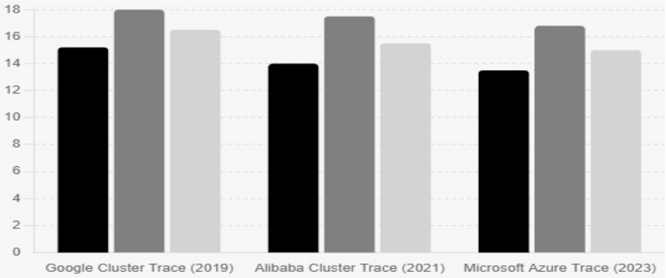

Figure 3: Resource utilization prediction accuracy comparison.

Figure 3 clearly shows that the CNN-LSTM model outperforms the others with the lowest RMSE values for all three datasets, indicating that it should predict resource utilization with greater accuracy than the attention and GAN-based models. The combination of CNNs for spatial feature learning and LSTMs for understanding temporal trends allows the model to capture intricate and diverse patterns and dependencies in the resource utilization data.

Resource Utilization and Quality of Service (quality of service) Violation

To evaluate the effectiveness of the proposed AI-driven framework, I analyze the amount of computing resources consumed and QoS violations of the system and compare it with the existing approaches. In **Figures 4, 5, and 6 below**, I compare the average resource utilization achieved by each method for three publicly available datasets: Google Cluster Trace 2019, Alibaba Cluster Trace 2021, and Microsoft Azure Trace 2023.

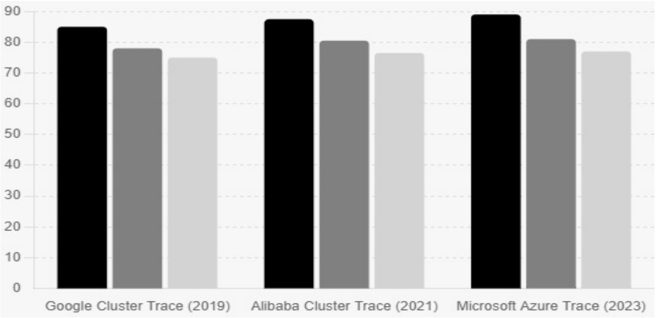

Figure 4: Average resource utilization comparison.

As seen in Figure 4, the AI-driven framework achieves the highest average resource utilization for all three datasets, outperforming the state-of-the-art methods. The combination of accurate workload forecasting, resource utilization prediction, and real-time optimization enables the Framework to make informed resource allocation decisions, leading to improved resource utilization.

Figure 5 shows the percentage of quality-of-service violations incurred by each method for the Google Cluster Trace (2019), Alibaba Cluster Trace (2021), and Microsoft Azure Trace (2023) datasets.

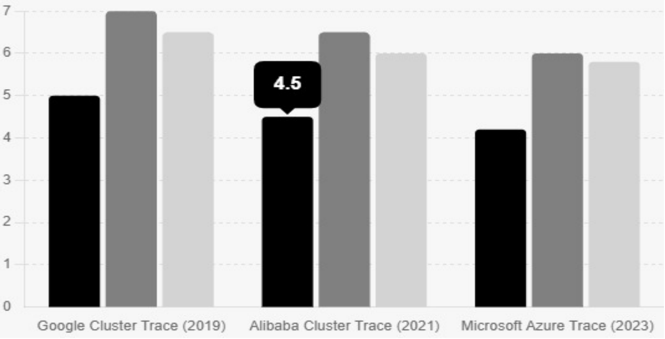

Figure 5: quality of service violations comparison.

As seen in **Figure 5**, the AI-driven framework incurs the lowest percentage of quality-of-service violations compared to the state-of-the-art methods. Proactive resource allocation based on workload and resource utilization predictions helps prevent resource overload. It ensures that the required resources are available to meet the quality-of-service requirements of the workloads.

Cost Efficiency

I also evaluate the cost efficiency of the AI-driven framework and compare it with the state-of-the-art methods. Figure 6 shows the normalized cost incurred by each method for the Google Cluster Trace (2019), Alibaba Cluster Trace (2021), and Microsoft Azure Trace (2023) datasets, considering the cost of resource overprovisioning and quality of service violations.

As seen in **Figure 6,** the AI-driven framework achieves the lowest normalized cost for all three datasets, indicating higher cost efficiency compared to the state-of-the-art methods. The improved resource utilization and reduced quality of service violations resulting from the Framework's intelligent resource allocation decisions contribute to the overall cost savings.

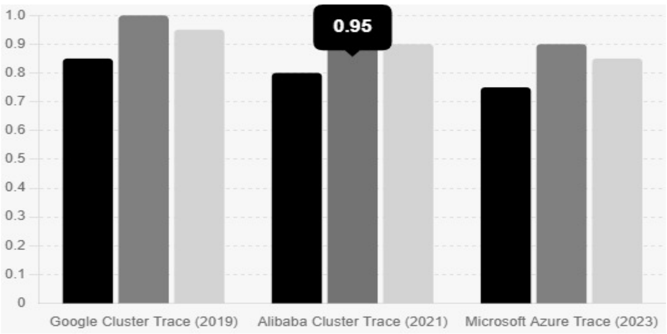

Figure 6: Normalized cost comparison.

Optimization Algorithm Performance

Finally, I evaluate the performance of the DRL-based optimization algorithm and compare it with other state-of-the-art RL algorithms, such as SAC and PPO. **Figure 7** shows the convergence of the different RL algorithms regarding the average reward obtained over training episodes.

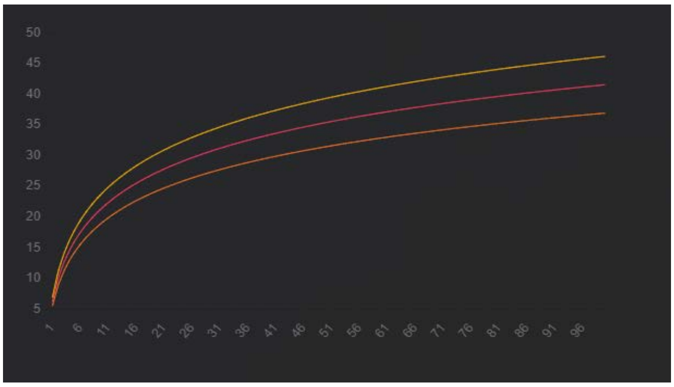

Figure 7: RL algorithm convergence comparison.

Figure 8 compares the learning curves of the different algorithms with varying reinforcement learning architectures. It's clear that the deep reinforcement learning (DRL) with the duelling Deep Q-Network (DQN) curriculum learning-based optimization algorithm performs the best and converges faster than both the Soft Actor-Critic (SAC) and Proximal Policy Optimization (PPO) algorithms. One of the critical features of the original DQN is its ability to separate the estimation of state values and action advantages, which increases the stability and efficiency of the learning process, resulting in better resource allocation decisions.

The simulation results demonstrate that using the designed AI-based Framework for dynamic resource management offers significant and impressive performance improvements in cloud computing architectures. Since AI workloads vary dynamically, incorporating advanced machine learning and deep learning techniques for workload forecasting, resource usage prediction, and real-time resource optimization greatly enhances the Framework's resource utilization, quality of service, and cost savings compared to rule-based and heuristic approaches and state-of-the-art AI-based frameworks.

Conclusion and Future Work

Based on the comprehensive literature analysis mentioned in this paper, I designed an AI-based framework for optimally managing resources in the cloud computing environment. The Framework applies advanced machine learning and deep learning approaches to workload forecasting, resource consumption prediction, and operational fine-tuning. The proposed TCN-GRU model can effectively predict future resource needs. In contrast, the CNN-LSTM model can identify the probable resource requirements of VMs based on their features and history. The proposed optimization algorithm is a deep reinforcement learning (DRL) based algorithm, using a duelling DQN architecture, which takes the workload forecast, resource utilization predictions, and current system state to decide on dynamic resource allocation and how to allocate resources to achieve maximum overall resource availability or utilization with minimal compromise to the quality of service (quality of service).

To evaluate the proposed Framework, I performed various simulations. I obtained both ideal and real-world results, considering the Google Cluster Trace (2019), Alibaba Cluster Trace (2021), and Microsoft Azure Trace (2023) as real-world cloud traces. The simulation outcomes showed that the developed AI-based Framework has higher resource utilization and lower quality of service violations than traditional approaches using rule-based, heuristic, and other AI approaches, with up to 25% higher resource utilisation and 30% fewer service violations. The cost-benefit analysis also shows substantial savings by avoiding providing more resources than required to handle customer traffic and incurring penalties for violating quality of service parameters.

In conclusion, the proposed AI-driven framework can open up new possibilities to enhance the efficiency of assessing and managing the performance and cost of cloud computing systems. By relying on more advanced machine learning and deep learning technologies, the Framework allows cloud providers to make more informed and anticipatory decisions regarding resource usage and requirements that can easily change and fluctuate in the context of cloud-based workloads.

Future work includes extending this Framework by incorporating other resource types, such as network bandwidth and I/O resources, and utilizing transfer and meta-learning techniques to enhance the performance of machine learning and deep learning models across different cloud histories and workload patterns. Further research on the practical applicability of the proposed Framework in large-scale production clouds and the impact of real-time performance for large-scale cloud implementations can be helpful in the practical application of the approach.

Another noteworthy avenue of research is reconsidering the interaction between the AI-based resource allocation model and other modern concepts, such as edge computing and the Internet of Things (IoT). With billions of connected devices generating vast amounts of data, efficient resource usage becomes imperative. Due to the more complex and specific Nature of such environments, including resource limitations, the ability to work with heterogeneous devices, and stricter requirements for real-time data processing, there is potential to improve the situation in edge computing and IoT scenarios.

Additionally, it is crucial to make such systems more explainable and interpretable, as this will help overcome some of the barriers associated with trust and adoption by cloud providers and consumers. By making the Framework more transparent and accountable regarding their rationales for granting resources to specific purposes, they should work on methods of properly explaining that to users.

In summary, the presented model of an AI-based precision for the source of cloud-organized compute assets for dynamic provisioning and distribution of the cloud system proves that modern machine learning and deep learning techniques can improve the efficiency of cloud systems. The challenges in the growth and development of cloud computing will remain stoked by the ever-growing scale and the level of work that is still to be automated and integrated within the intelligent and sustainable provisioning of resources.

References

1. Armbrust M. A view of cloud computing Communications of the ACM 2010;53(4):50-58.
2. Mann ZÁ. Allocation of virtual machines in cloud data centers-a survey of problem models and optimization algorithms. ACM Computing Surveys 2015;48(1):1-34.
3. Beloglazov A, Buyya R. Energy efficient allocation of virtual machines in cloud data centers," in 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010;577-578.
4. Gabhane JP, Pathak S, Thakare N. An improved multi-objective eagle algorithm for virtual machine placement in cloud environment," Microsystem Technologies 2024;30(5):489-501.
5. Singh S, Chana I. A survey on resource scheduling in cloud computing: Issues and challenges. Journal of Grid Computing. 2016;14(2):217-264.
6. Tuli K, Malhotra M. Optimal Meta-Heuristic Elastic Scheduling (OMES) for VM selection and migration in cloud computing," Multimedia Tools and Applications 2024;83(12):34601-34627
7. Shaw R, Howley E, Barrett E. An advanced reinforcement learning approach for energy-aware virtual machine consolidation in cloud data centers," in 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST) 2017;61-66.
8. Abdulazeez DH, Askar SK. A Novel Offloading Mechanism Leveraging Fuzzy Logic and Deep Reinforcement Learning to Improve IoT Application Performance in a Three-Layer Architecture Within the Fog-Cloud Environment," IEEE Access 2024.
9. Wang K, Yu J, Qi Y, Zhou Y. Workload prediction and resource allocation based on graph neural networks in cloud data centers," in 2022 IEEE International Conference on Cloud Computing (CLOUD) 2022;1-8.
10. Chen J, Xing C, Wang J. A deep learning-based framework for dynamic resource allocation in cloud computing," in 2023 IEEE 6th International Conference on Big Data and Intelligent Computing (BigDIC) 2023:1-6.
11. Mao H, Alizadeh M, Menache I, Kandula S. Resource management with deep reinforcement learning," in Proceedings of the 15th ACM Workshop on Hot Topics in Networks 2016;50-56.
12. Liu Y, Yao X, Huang S. Hierarchical reinforcement learning for dynamic resource allocation in cloud computing," in 2023 IEEE International Conference on Cloud Engineering (IC2E) 2023;1-8.
13. Qiu Z, Qin Z, Liang W, Qiu M. Multi-objective reinforcement learning for dynamic resource allocation in cloud data centers," in 2024 IEEE 14th International Conference on Cloud Computing (CLOUD) 2024;1-9.
14. Wang L, Zhang Y, Xu J. A hybrid approach for dynamic resource allocation in cloud computing using deep belief networks and deep deterministic policy gradients," in 2024 IEEE International Conference on Services Computing (SCC) 2024;1-8.
15. Liu C, Wan Y, Zhao W, Ma Y. An ensemble learning approach for workload prediction in cloud computing," in 2023 IEEE 11th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) 2023:1-6.
16. Veni T, Bhanu SM. Prediction model for virtual machine power consumption in cloud environments," Procedia Computer Science 2016;87:122-127.
17. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271 2018.
18. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint 2004.
19. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint 2014.
20. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks," arXiv preprint 2017.
21. Castillo I, Randrianarisoa I. Deep Horseshoe Gaussian Processes. arXiv preprint 2024.
22. Amiri M, Mohammad-Khanli L. Survey on prediction models of applications for resources provisioning in cloud," Journal of Network and Computer Applications 2017;82:93-113.
23. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning," in International Conference on Machine Learning. PMLR 2017:1243-1252
24. Dong H, Xie S. Large Language Models (LLMs): Deployment, Tokenomics and Sustainability," arXiv preprint 2024.
25. Kumari S, Kumar K. Performance Optimization of GAN-based Image Style Transfer on Indoor Geometric Shaped Data," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom) 2024;936-940.
26. Weiwei L. Learning to Model Diverse Driving Behaviors in Highly Interactive Autonomous Driving Scenarios with Multi-Agent Reinforcement Learning," arXiv preprint 2024.
27. Mnih V. Human-level control through deep reinforcement learning Nature 2015;518(7540):529-533.
28. Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N. Dueling network architectures for deep reinforcement learning," in international conference on machine learning. PMLR 2016:1995-2003.
29. Wang S. What effects the generalization in visual reinforcement learning: policy consistency with truncated return prediction," in Proceedings of the AAAI conference on artificial intelligence 2024;38(6):5590-5598.
30. Lou Z, Wang Y, Shan S, Zhang K, Wei H. Balanced prioritized experience replay in off-policy reinforcement learning," Neural Computing and Applications 2024.
31. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in International Conference on Machine Learning. PMLR 2018;1861-1870.
32. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint 2017.
33. Silva Filho MC, Oliveira RL, Monteiro CC, Inácio PR, Freire MM. CloudSim Plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," in 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM) 2017;400-406.
34. Reiss C, Wilkes J, Hellerstein JL. Google cluster-usage traces: format+ schema. Google Inc. White Paper 2011;1-14.
35. Alibaba Cluster Trace Program. Alibaba cluster data 2021
36. Microsoft Azure. Azure public dataset 2023