Journal of Advanced Computing Systems (JACS)
www.scipublication.com

AI-Driven Cloud Resource Optimization Framework for Real-Time Allocation

Vijay Ramamoorthi
Independent Researcher

DOI: 10.69987/JACS.2021.10102

Keywords

Artificial intelligence
Cloud Resource
Cost Efficiency

Abstract

Cloud resource optimization is a critical challenge in modern distributed systems, particularly as workloads become more dynamic and unpredictable. This paper presents an AI-driven framework for cloud resource optimization that autonomously manages and allocates resources in real time based on workload demands. By leveraging machine learning techniques, including predictive modeling and reinforcement learning, the framework dynamically adjusts CPU, memory, and bandwidth allocations to improve both cost efficiency and system performance. The proposed framework continuously learns from historical and real-time data, enabling proactive resource scaling to prevent over- or under-provisioning. We evaluate the framework using historical resource usage data from leading cloud platforms, demonstrating significant improvements in resource utilization, cost savings, and application performance. The results highlight the framework's effectiveness in optimizing cloud environments for AI-driven applications and web services. Future research could expand this approach to hybrid cloud-edge environments and further generalize predictive models to adapt across diverse workloads and cloud platforms.

Introduction

As the reliance on cloud computing continues to expand across industries, the effective management and optimization of cloud resources have become critical challenges. Modern cloud environments host a variety of applications, from simple web services to complex artificial intelligence (AI) workloads, all of which require dynamic and efficient resource allocation [1]. With the rise of large-scale, distributed systems, ensuring that resources such as CPU, memory, and bandwidth are efficiently utilized has become a pressing concern for cloud service providers and organizations alike. Inefficient resource management can lead to overprovisioning, increased operational costs, and performance degradation, while under-provisioning can lead to system bottlenecks and failures, negatively affecting the user experience and application performance [2]–[4].

Cloud resource optimization is further complicated by the unpredictable nature of workloads in these environments [5], [6]. Workload demands can fluctuate widely, making it difficult to manually allocate resources in a way that balances performance with cost-efficiency. This challenge is particularly acute in AI-driven applications, which often require intensive computational power for tasks such as model training, real-time inference, and data processing. Traditional resource management techniques, which rely on static rules or basic heuristics, are ill-equipped to cope with the complexity and dynamic nature of such workloads. The ability to predict future resource requirements and adjust allocations in real time is essential to optimizing cloud environments in a way that ensures high performance while minimizing costs [7].

In response to these challenges, this paper proposes the development of an AI-driven cloud-based framework that autonomously manages and optimizes cloud resources in real-time. The proposed framework leverages predictive modeling and advanced optimization techniques to dynamically allocate resources based on both current and anticipated workload demands. By integrating AI techniques such as machine learning and reinforcement learning, the framework can forecast future resource needs based on historical usage patterns and real-time data. This predictive capability allows the cloud environment to proactively scale resources up or down, improving efficiency and reducing the risks associated with resource overutilization or underutilization.

The contribution of this study lies in its development of a novel AI-driven framework for real-time cloud resource optimization. The framework integrates machine learning techniques, including predictive modeling and reinforcement learning, to dynamically manage the allocation of cloud resources based on current and anticipated workload demands. By enabling autonomous resource scaling, this approach enhances both cost efficiency and system performance in cloud environments. Unlike traditional static methods, the proposed framework proactively adjusts resources to optimize utilization, ensuring cloud applications, particularly AI-driven workloads, meet performance requirements while avoiding over- or under-provisioning. This work also offers empirical insights through performance evaluations using historical data from major cloud platforms, demonstrating significant improvements in resource utilization, cost savings, and system performance. Additionally, this research opens avenues for future work, including the exploration of hybrid cloud-edge environments and the development of generalized predictive models applicable across various cloud platforms.

Literature Review

The management of cloud resources is a critical challenge due to the dynamic and unpredictable nature of workloads, particularly in AI-driven applications. Several studies have addressed the need for efficient resource management systems that can adapt to real-time demands. Ilager et al. (2020) proposed an AI-centric approach to manage cloud data centers, highlighting the feasibility of AI-driven solutions for resource optimization in large-scale distributed systems [8]. Another approach by Kandan and Manimegalai (2019) focused on enhancing Quality of Service (QoS) in cloud environments using multi-agent-based dynamic resource allocation techniques [9]. Martino et al. (2019) explored AI-powered tools for optimizing cloud performance and anomaly detection, emphasizing the need for a common cloud service representation to support optimization [10]. A hybrid optimization algorithm combining teaching-learning and grey wolves optimization to balance load distribution across virtual machines in the cloud is presented in [11].

The role of virtualization technology in optimizing cloud resource allocation was emphasized in some literature who introduced the concept of skewness to enhance server utilization and save energy [12], [13]. Ghanbari et al. (2012) focused on feedback-based optimization in private clouds, which dynamically updated performance models to improve cost efficiency [14]. A genetic algorithm for QoS-based resource allocation, showcasing its effectiveness in meeting budget constraints while optimizing resource utilization is studied in [15]. Dai et al. (2018) developed a multi-objective optimization algorithm to manage cloud infrastructure for big data applications, demonstrating its ability to outperform traditional approaches by 20% [16]. Gai et al. (2018) addressed the resource management challenges in cyber-physical systems by proposing algorithms that minimized bottlenecks in heterogeneous cloud environments [17]. Islam et al. (2012) used prediction-based strategies to improve adaptive resource provisioning, demonstrating the benefits of machine learning in cloud environments [18].

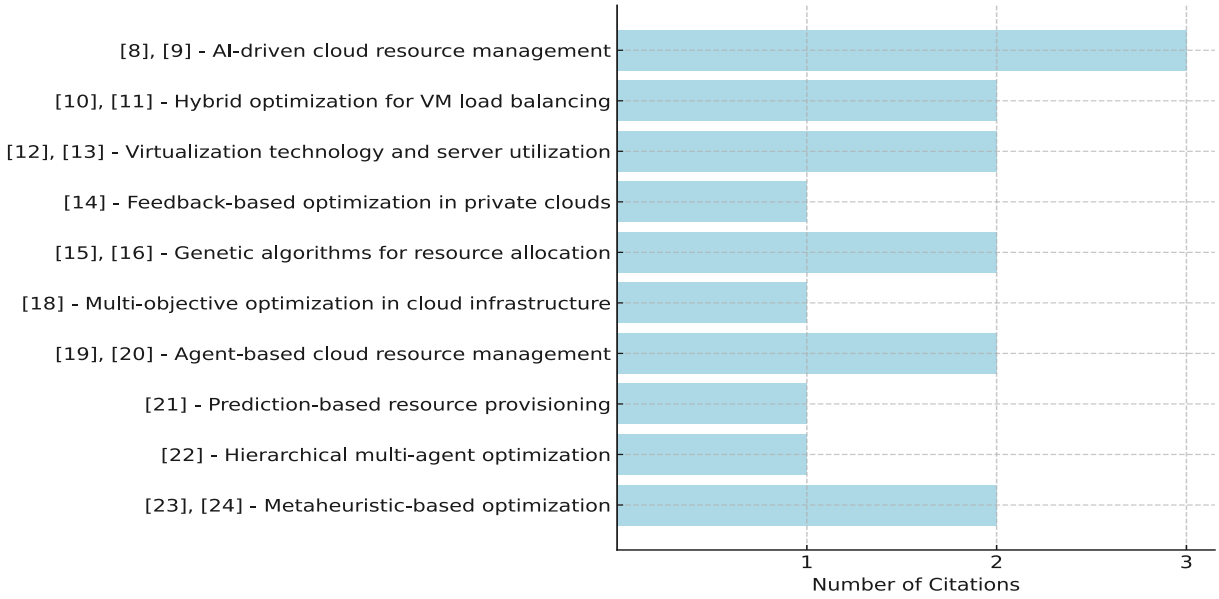


Figure 1 Summary of the literature review

Simic et al. (2019) extended the WoBinGO framework for large-scale optimization in cloud computing, achieving cost savings and faster delivery of optimization results [19]. Gao et al. (2020) proposed a hierarchical multi-agent optimization algorithm for resource allocation, combining genetic algorithms and multi-agent optimization to enhance resource utilization and reduce costs [20]. Papagianni et al. (2013) formulated a mixed-integer programming model for optimal resource allocation in networked clouds, emphasizing the importance of joint optimization of computing and networking resources [21]. Singh and Goraya (2019) presented a hybrid optimization scheme for dynamic resource management using fruit fly optimization, which significantly reduced energy consumption and SLA violations [22]. Chaabouni et al. (2013) proposed an agent-based model for cloud resource management, providing cloud service providers with an automated tool for optimizing resource allocation [23]. Akintoye and Bagula (2017) optimized virtual resource allocation in cloud environments using the Hungarian Algorithm to minimize job execution time [24]. Sun et al. (2013) developed a wind-driven optimization algorithm for resource allocation in distributed clouds, demonstrating the feasibility of using non-price attributes to enhance cloud performance [25]. Strumberger et al. (2019) introduced a hybrid whale optimization algorithm for resource scheduling, improving cloud performance compared to existing metaheuristics [26]. Lastly, Sun et al. (2016) presented the ROAR framework, which automated resource allocation for web applications by optimizing cost and performance in cloud environments [27].

While existing literature has made considerable progress in addressing cloud resource optimization, key challenges remain in adapting these frameworks to hybrid cloud-edge environments, developing generalized predictive models, balancing cost and performance, and scaling optimization techniques to handle large-scale distributed systems. These gaps provide opportunities for further exploration and innovation in the design of AI-driven cloud resource management frameworks.

Framework Design

The proposed AI-driven cloud resource optimization framework consists of multiple components that work together to manage and optimize the allocation of cloud resources. This section describes the key design elements, including the data ingestion pipeline, AI modeling for resource prediction, and optimization algorithms that dynamically adjust resource allocation. The framework's flexibility allows it to be applied to a variety of use cases, such as web services and AI workloads, ensuring improved performance and cost efficiency.

Cloud Resource Management

Cloud resource management is the core function of the framework, responsible for monitoring resource usage and making real-time adjustments based on AI-driven predictions. The framework is designed to autonomously allocate resources such as CPU, memory, and bandwidth, ensuring that the cloud infrastructure can meet the dynamic needs of different applications without over-provisioning or under-provisioning resources. The first step in the resource management process is data ingestion, which involves continuously monitoring the cloud environment for resource utilization data. This data is collected from various running applications and services, providing information about CPU usage, memory consumption, bandwidth utilization, and network traffic patterns. The data ingestion pipeline is critical for ensuring that the AI model has access to both historical and real-time data, allowing it to make accurate predictions about future resource demands. In practice, data is ingested from cloud monitoring tools and application logs, which feed into the system at regular intervals.

At the heart of the framework is an AI model designed to forecast future resource requirements based on historical usage patterns and real-time data. The model leverages machine learning algorithms, such as time-series analysis and deep learning techniques, to understand resource demand trends. By training on historical cloud resource data, the model is able to predict how much CPU, memory, and bandwidth will be required for various workloads, particularly during peak demand periods or under fluctuating traffic conditions. A reinforcement learning-based approach can also be integrated into the AI model, enabling it to continuously improve its predictions by learning from the outcomes of previous resource allocation decisions. This adaptive learning process ensures that the AI model remains responsive to changing workloads and evolving application requirements.

Optimization Algorithm: Reinforcement Learning Approach

In the context of cloud optimization, Reinforcement Learning (RL) operates by interacting with the cloud infrastructure to manage resources dynamically. The **state** represents the current condition of the cloud system, including metrics such as CPU usage, memory consumption, bandwidth, and overall performance (e.g., response time). Based on the state, the RL **agent** takes **actions**, such as adjusting the number of CPUs, modifying memory allocation, or scaling virtual machines. After each action, the agent receives

rewards, which reflect the effectiveness of the decision. If the action improves performance and reduces costs, the reward is positive. However, if the action leads to wasted resources or degraded performance, the reward is negative. Over time, the agent develops a **policy**—a strategy to select actions that maximize long-term rewards. As the RL agent continues interacting with the environment, it balances exploration of new strategies and exploitation of known effective actions, allowing it to adapt and optimize resource allocation based on real-time performance and cost feedback. A simplified functions are shown in Figure 2.

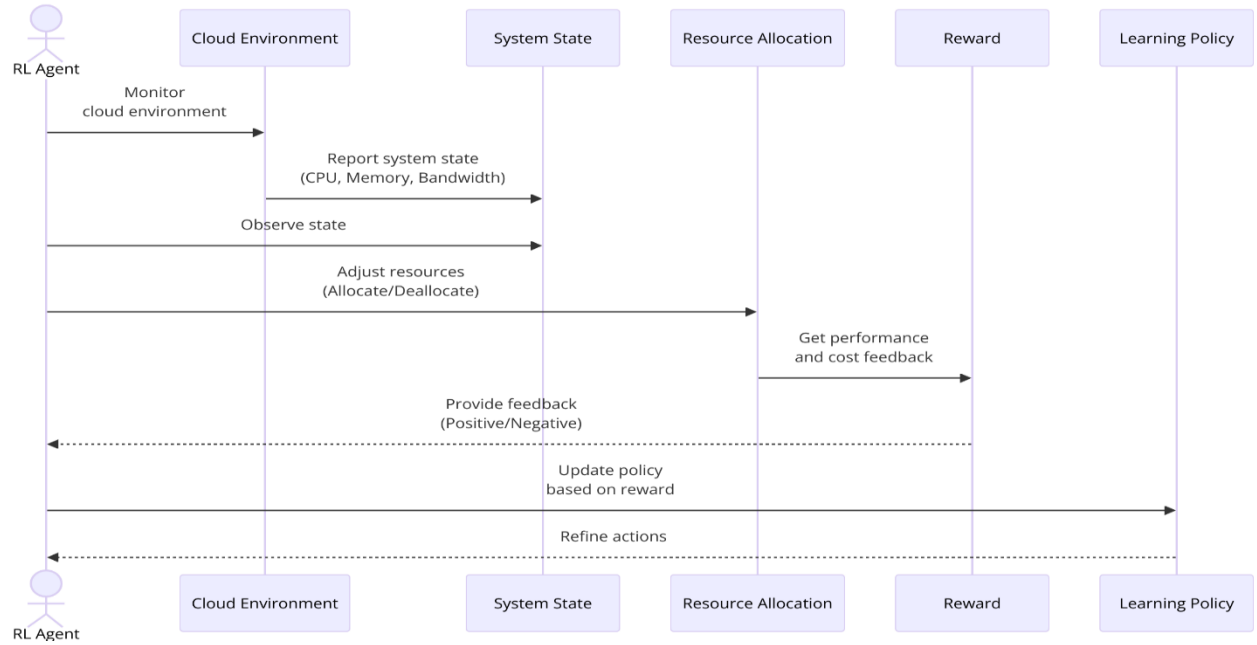


Figure 2 The sequence diagram illustrates how a Reinforcement Learning (RL) agent interacts with the cloud environment for dynamic resource optimization. The agent observes the system's state, takes actions to allocate or deallocate resources, and receives feedback in the form of rewards. The reward informs the agent about the efficiency of the action, helping it refine its strategy (policy) over time. This continuous interaction allows the system to adapt to changing workloads, maximizing performance and minimizing costs.

Learning from Experience: Continuous Adaptation

As the RL agent takes actions, it learns from the outcomes and updates its policy. The agent balances **exploration** (trying new actions to discover better strategies) and **exploitation** (using known actions that have worked in the past). This process allows the system to continuously adapt to changing workloads and resource demands.

The reward function plays a crucial role in guiding the RL agent's behavior. It needs to balance the following objectives:

- **Cost Efficiency:** Minimize resource usage while keeping costs low.
- **Performance Optimization:** Ensure the application meets performance standards, such as low response times and high throughput.
- **Service-Level Agreement (SLA) Compliance:** Ensure resources are allocated to avoid SLA violations (e.g., exceeding latency limits).

For example, when a web service experiences a traffic spike, the RL agent can proactively allocate additional resources to handle the increased load. If the traffic subsides, the agent learns to scale down resources, preventing over-provisioning and minimizing costs.

Analysis and Results

In this section, we present a comprehensive analysis of the AI-driven cloud resource optimization framework, based on the key metrics of resource utilization, cost savings, and performance. The analysis uses historical cloud resource usage data from leading cloud platforms such as AWS and Google Cloud to train predictive models. The evaluation focuses on the effectiveness of dynamic resource allocation in optimizing resource usage, minimizing costs, and maintaining high performance for cloud applications.

The dataset used for model training consists of historical resource usage data collected from cloud environments. This data includes time-series metrics related to CPU utilization, memory consumption, and network

bandwidth usage across a variety of applications and workloads. The workloads range from simple web services to more computationally demanding AI tasks such as model training and real-time inference. The dataset captures varying patterns of workload behavior over time, including peak demand periods, fluctuations in traffic, and periods of low activity. This historical data is critical for training the AI model to recognize patterns and forecast future resource demands, allowing for proactive scaling of cloud resources. By leveraging this dataset, the framework is designed to predict resource utilization patterns in real time and make adjustments accordingly. This ability to anticipate future demand ensures that cloud resources are allocated more efficiently, reducing waste and improving overall system performance. The predictive model uses machine learning techniques to identify relationships between past usage data and future resource needs, enabling the cloud environment to operate more intelligently. To evaluate the performance of the AI-driven framework, three key metrics were used: resource utilization, cost savings, and application performance. These metrics provide insight into how well the system manages cloud resources compared to traditional non-optimized environments.

Resource Utilization

Resource utilization is a critical factor in determining the efficiency of cloud environments. Poor resource management can lead to either over-provisioning, where too many resources are allocated, or under-provisioning, where insufficient resources are available to meet demand. Both scenarios negatively impact performance and operational costs. In the optimized cloud environment, the AI-driven framework was able to dynamically allocate resources, such as CPU, memory, and network bandwidth, based on predicted workload demand. The model continuously monitors usage patterns and adjusts allocations in real time to match current and future requirements. As a result, resource utilization in the optimized environment showed significant improvements when compared to the non-optimized environment.

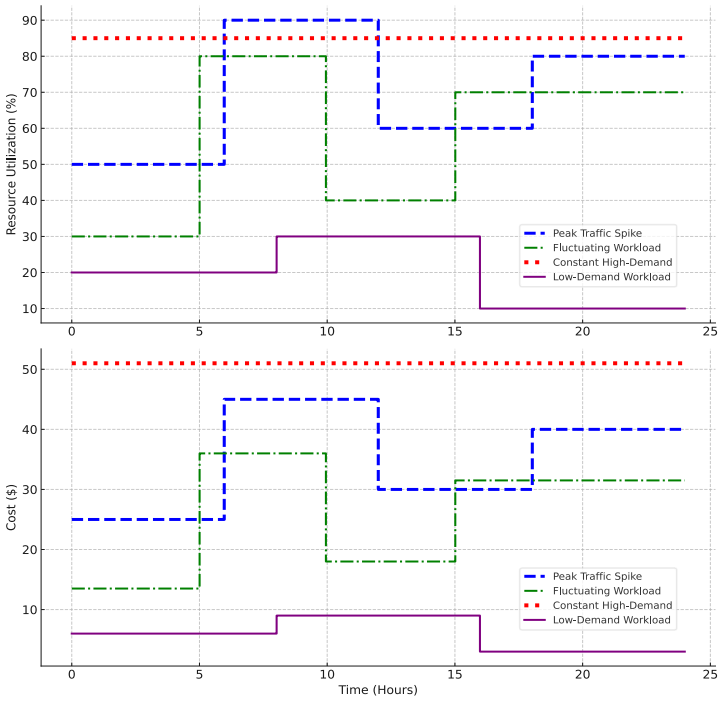


Figure 3 Performance of four workload scenarios—Peak Traffic Spike, Fluctuating Workload, Constant High-Demand, and Low-Demand—by evaluating resource utilization (top panel) and cost efficiency (bottom panel) over time. Resource utilization is represented as a percentage, showing how efficiently resources like CPU and memory are allocated in response to each workload. The cost plot illustrates the associated expenses based on the dynamic resource provisioning. The AI-driven framework optimizes resource allocation, reducing costs during periods of fluctuating and low demand while maintaining efficiency during peak loads.

Optimized Environment: In this environment, the AI model proactively adjusts resources, ensuring that only the necessary amount of CPU, memory, and bandwidth is allocated at any given time. This leads to better resource utilization during both peak and low-demand periods. For instance, during times of low traffic, the system scales down unused resources, thereby preventing unnecessary over-provisioning. Conversely,

during high-demand periods, such as AI model training or web service traffic spikes, the model quickly scales up resources to meet demand, avoiding performance bottlenecks.

Non-Optimized Environment: In contrast, the non-optimized environment relies on static resource allocation policies that do not adapt to changing workload conditions. Resources are often over-provisioned to ensure adequate performance during peak traffic, but this results in underutilized resources during periods of low activity. Furthermore, during unexpected surges in demand, the system may face under-provisioning, leading to performance issues such as increased response times and higher error rates.

Overall, the AI-driven framework demonstrated a significant improvement in resource utilization by eliminating the inefficiencies associated with static resource management. By dynamically scaling resources in response to workload demands, the system ensures optimal resource usage at all times. Cost savings and resource utilization is hown in Figure 3.

Cost Savings

Cost efficiency is a major driver for cloud optimization efforts, as cloud resources are often billed based on usage. The AI-driven framework was designed to reduce operational costs by minimizing unnecessary resource allocation while maintaining performance standards. This was achieved through the model's ability to predict future demand and adjust resources in real time, thus preventing the over-provisioning of expensive cloud resources.

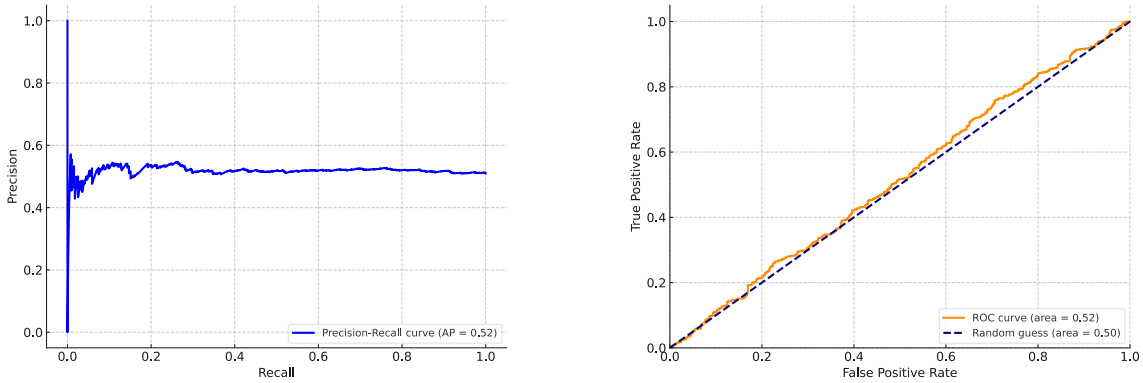
Optimized Environment: In the optimized environment, the AI system continuously monitors and predicts workload trends, scaling resources up or down to align with actual usage. This real-time adjustment results in substantial cost savings, particularly during periods of low demand. For example, when an AI training workload is not running, the system scales down the number of allocated virtual machines (VMs) or container instances, thus reducing costs. Similarly, during high-demand periods, the system scales resources just enough to handle the load without over-allocating, preventing wasteful spending.

Non-Optimized Environment: In contrast, the non-optimized environment does not have the capability to adjust resources dynamically. As a result, the system typically allocates more resources than necessary to ensure performance during high-demand periods. This leads to significant resource wastage during off-peak times, increasing overall cloud costs. Additionally, manual adjustments to resource allocations are often reactive and may result in temporary under-provisioning, which can incur penalties or lead to performance degradation that affects service-level agreements (SLAs).

The cost savings realized by the AI-driven framework were a direct result of its ability to optimize resource allocation in response to real-time demand forecasts. The framework minimized resource wastage and ensured that cloud resources were used efficiently, leading to a reduction in overall cloud spending.

Performance

Application performance is a critical factor in cloud environments, especially for applications that require high computational power, such as AI model training and real-time inference. The AI-driven framework was evaluated based on key performance metrics, including response time and throughput, to determine how well it handled dynamic resource allocation. The performance parameters is shown in Figure 4.



a. ROC curve, comparing the model's true positive rate (TPR) against its false positive rate (FPR) across different thresholds

b. Precision-Recall curve, with an average precision (AP) score of 0.52

Figure 4 Evaluation of Resource Provisioning Prediction using ROC and Precision-Recal

Optimized Environment: In the optimized environment, the AI framework maintained high performance by adjusting resources in real time based on workload demands. During peak traffic periods, the system scaled resources to ensure that response times remained low and throughput remained high. This was particularly important for AI applications that require significant compute power for tasks such as model training. The dynamic scaling ensured that enough resources were available to handle the load, while preventing over-provisioning during periods of low activity.

Non-Optimized Environment: In the non-optimized environment, static resource allocation led to performance degradation during periods of high demand. Without the ability to dynamically adjust resources, the system struggled to maintain low response times and high throughput when workloads spiked. This negatively impacted application performance, particularly for real-time applications where latency is critical.

The AI-driven framework demonstrated a clear advantage in maintaining high performance across different workloads. By dynamically adjusting resource allocations, the system ensured that applications met their performance requirements without incurring unnecessary costs.

Conclusion and Future Work

In this paper, we proposed an AI-driven cloud-based framework for dynamic resource optimization, addressing the growing challenges of efficiently managing cloud resources in large-scale, distributed environments. By leveraging predictive modeling and reinforcement learning techniques, the framework autonomously manages the allocation of resources such as CPU, memory, and bandwidth, optimizing them in real-time based on workload demands. This approach not only improves resource utilization but also reduces operational costs by dynamically scaling resources according to predicted needs. The framework was demonstrated to effectively enhance both performance and cost-efficiency in a variety of use cases, including AI-driven applications and web services, which require responsive and scalable infrastructure.

The performance evaluation showed significant improvements in key metrics, including resource utilization, cost savings, and application performance. The AI-driven framework's ability to predict future resource demands and make proactive adjustments to resource allocation led to a substantial reduction in unnecessary provisioning, preventing both overuse and under-provisioning. This resulted in optimized cloud environments that maintained high performance levels even during periods of fluctuating workload demands.

Future Work

While the proposed framework offers promising results, several avenues for future research remain open. One potential area for exploration is the extension of the framework to hybrid cloud environments, where cloud resources are distributed across both cloud and edge infrastructures. The integration of edge computing into cloud optimization presents unique challenges due to the diverse and distributed nature of edge nodes. Future research could investigate how AI-driven resource management techniques can be adapted to optimize resources across hybrid cloud-edge architectures, especially for low-latency applications. Additionally, the development of more generalized predictive models for resource forecasting remains an important area for further investigation. While the current framework demonstrated effectiveness using specific datasets, creating models that can generalize across different cloud platforms and workload types would enhance the framework's applicability to broader scenarios.

References

[1] P. Hofmann, "The Limits of Public Clouds for Business Applications --An overly simplistic reliance on the utility model risks blinding us to the real opportunities and challenges of cloud computing," in *Proceedings of the 2010 International Conference on E-Business Intelligence*, China, 2010.

[2] Kanniga Devi R., M. Gurusamy, and Vijayakumar P., "An efficient cloud data center allocation to the source of requests," *J. Organ. End User Comput.*, vol. 32, no. 3, pp. 23–36, Jul. 2020.

[3] N. M. Gonzalez, T. C. M. de B. Carvalho, and C. C. Miers, "Cloud resource management: towards efficient execution of large-scale scientific applications and workflows on complex infrastructures," *J. Cloud Comput. Adv. Syst. Appl.*, vol. 6, no. 1, Dec. 2017.

[4] M. A. Alworafi, A. Dhari, S. A. E. Booz, and S. Mallappa, "Budget-aware task scheduling technique for efficient management of cloud resources," *Int. J. High Perform. Comput. Netw.*, vol. 14, no. 4, p. 453, 2019.

[5] J. M. Luna, C. T. Abdallah, and G. L. Heileman, "Probabilistic optimization of resource distribution and encryption for data storage in the cloud," *IEEE Trans. Cloud Comput.*, vol. 6, no. 2, pp. 428–439, Apr. 2018.

[6] A. Al-Mansoori, J. Abawajy, and M. Chowdhury, "SDN enabled BDSP in public cloud for resource optimization," *Wirel. Netw.*, Nov. 2018.

[7] F. Nzanywayingoma and Y. Yang, "A literature survey on resource management techniques, issues and challenges in cloud computing," *TELKOMNIKA*, vol. 15, no. 4, p. 1918, Dec. 2017.

[8] S. Ilager, R. Muralidhar, and R. Buyya, "Artificial Intelligence (AI)-centric management of resources in modern Distributed Computing Systems," *arXiv [cs.DC]*, 09-Jun-2020.

[9] M. Kandan and R. Manimegalai, "Optimum resource allocation techniques for enhancing quality of service parameters in cloud environment," in *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*, Cham: Springer International Publishing, 2020, pp. 831–839.

[10] B. D. Martino, A. Esposito, and E. Damiani, "Towards AI-powered multiple cloud management," *IEEE Internet Comput.*, vol. 23, no. 1, pp. 64–71, Jan. 2019.

[11] J. Agarkhed and R. Ashalatha, "Dynamic resource allocation mechanism using SLA in cloud computing," in *Advances in Intelligent Systems and Computing*, Singapore: Springer Singapore, 2017, pp. 731–740.

[12] L. S. Subhash and D. K. P. Thooyamani, "Allocation of resource dynamically in cloud computing environment using virtual machines," *Int. J. Adv. Technol.*, vol. 08, no. 04, 2017.

[13] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, pp. 1107–1117, Jun. 2013.

[14] H. Ghanbari, B. Simmons, M. Litoiu, and G. Iszlai, "Feedback-based optimization of a private cloud," *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 104–111, Jan. 2012.

[15] C. Xu, B. Chen, P. Fu, and H. Qian, "A dynamic resource allocation model for guaranteeing quality of service in software defined networking based cloud computing environment," in *Cloud Computing and Security*, Cham: Springer International Publishing, 2015, pp. 206–217.

[16] W. Dai, L. Qiu, A. Wu, and M. Qiu, "Cloud infrastructure resource allocation for big data applications," *IEEE Trans. Big Data*, vol. 4, no. 3, pp. 313–324, Sep. 2018.

[17] K. Gai, M. Qiu, H. Zhao, and X. Sun, "Resource management in sustainable cyber-physical systems using heterogeneous cloud computing," *IEEE Trans. Sustain. Comput.*, vol. 3, no. 2, pp. 60–72, Apr. 2018.

[18] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 155–162, Jan. 2012.

[19] V. Simic, B. Stojanovic, and M. Ivanovic, "Optimizing the performance of optimization in the cloud environment–An intelligent auto-scaling approach," *Future Gener. Comput. Syst.*, vol. 101, pp. 909–920, Dec. 2019.

[20] X. Gao, R. Liu, and A. Kaushik, "Hierarchical multi-agent optimization for resource allocation in cloud computing," *arXiv [cs.DC]*, 12-Jan-2020.

[21] C. Papagianni, A. Leivadeas, S. Papavassiliou, V. Maglaris, C. Cervello-Pastor, and A. Monje, "On the optimal allocation of virtual resources in cloud computing networks," *IEEE Trans. Comput.*, vol. 62, no. 6, pp. 1060–1071, Jun. 2013.

[22] J. Singh and M. S. Goraya, "Multi-objective hybrid optimization based dynamic resource management scheme for cloud computing environments," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2019.

[23] T. Chaabouni, H. Kchaou, and M. Khemakhem, "Agent technology based resources management in Cloud Computing," in *2013 World Congress on Computer and Information Technology (WCCIT)*, Sousse, Tunisia, 2013.

[24] S. B. Akintoye and A. Bagula, "Optimization of virtual resources allocation in cloud computing environment," in *2017 IEEE AFRICON*, Cape Town, 2017.

[25] J. Sun, X. Wang, M. Huang, and C. Gao, "A cloud resource allocation scheme based on microeconomics and wind driven optimization," in *2013 8th ChinaGrid Annual Conference*, Los Alamitos, CA, USA, 2013.

[26] I. Strumberger, N. Bacanin, M. Tuba, and E. Tuba, "Resource scheduling in cloud computing based on a hybridized whale optimization algorithm," *Appl. Sci. (Basel)*, vol. 9, no. 22, p. 4893, Nov. 2019.

[27] Y. Sun, J. White, S. Eade, and D. C. Schmidt, "ROAR: A QoS-oriented modeling framework for automated cloud resource allocation and optimization," *J. Syst. Softw.*, vol. 116, pp. 146–161, Jun. 2016.