Ai-driven Optimization of Cost, Performance, And Resource Allocation in Multi-Tenant Cloud Data Platforms

Rambabu Bandam[1], Senthil Raj Subramaniam[2]

[1]Director of Engineering, Data & Analytics, Nike Inc
[2]Information Technology Manager, Insurance(ADM), Cognizant Technology Solutions

Abstract

As cloud computing emerges as one of the influential characteristics in the IT structure of the modern world, cost, performance, and resource management in multi-tenant cloud data structures are increasingly vital for both cloud service providers and enterprises. Multi-tenant environments are typically already complicated due to the issues associated with shared resources because the environment must serve numerous clients with diverse needs and requirements efficiently. This paper provides a study of the issues in multi-tenant cloud platforms and presents an optimal AI-based solution to deal with these issues by reducing cost, enhancing performance metric and dynamic resource management.

The presented framework includes state-of-art AI methods and tools such as predictive analytics for the workload prediction, clustering for tenants' behavior analysis, and reinforcement learning for an agile and real-time resource allocation. These concepts are then used for anticipation of system workloads and the distribution of resources depending on the tenant's requirements and needs and the reduction of operational costs at the same time without an impact on efficiency. Also, it integrates a multicriteria optimization that has consideration of costs, performance and resources consumption in accordance with standards of specific tenants.

To illustrate the implementation of the proposed AI-driven optimization framework in a general multi-tenant cloud system for different application workloads, an example case is explained as follows. In this case, certain benefits of the proposed method include a notable decrease in the overall operating cost by 22 %, increase in efficiency by 35 %, and improper resource utilization compared to other optimization techniques. These could only confirm the efficiency of the AI strategies to manage cloud resources and further confirm that AI can make a huge difference in future cloud systems management.

Thus, this research also discusses the future of AI in enhancing cloud optimization with the identified areas for future work including federated learning in distributed cloud environment and application of explainable AI in enhancing cloud computing. In conclusion, this study adds to the existing literature concerning the application of AI in cloud computing as well as offering recommendations that CSPs may use to improve their resource management techniques with respect to scalability and flexibility.

Keywords: Ai-driven Cloud Optimization, Multi-tenant Cloud Platforms, Resource Allocation in Cloud Computing, Cost-performance Optimization, Reinforcement Learning in Cloud Systems

1. Introduction

1.1 Background on Cloud Computing and Multi-Tenancy

Cloud computing turned out to be one of the major revolutionary models within the last decade, relocating the approach to store and process information from local environments to remote ones. That is a model that enables users to access various computing resources including storage, processing, and networking among others without owning physical assets that provide the infrastructure. The cloud computing paradigm shift is based on the notion known as multi-tenancy that enables many separate customer, organization, or business division (referred to as tenants) to use the same physical resources while being informed that their data are kept separate.

Large-scale fully managed MPP platforms from AWS Redshift, Google BigQuery and Snowflake are nowadays part of the standard tool kit for new generation data-oriented industries. These platforms also bring the cost efficiency as the tenants share the underlying infrastructure and can gain new resources on demand instead of purchasing own servers. In addition, multi-tenancy leads to elasticity as the availability of the pertinent resources is based on the present usage by the tenants. That said, the nature and characteristics of multiple tenants create certain difficulties in achieving optimal performance, cost, and resource distribution among tenants. As more and more business organizations move to the cloud, it becomes a paramount to meet these to ensure cloud provider stay ahead in the market and provide services.

1.2 Challenges in Cost, Performance, and Resource Allocation

In this method, however, these potential benefits are headed by several problems, especially when it comes to instilling costs, performance, and resources. The multi-tenant environment is characterized by different activity levels between tenants hence the usage is dynamic. Therefore, any given client load or ramping up activities can influence the alterations in the other tenants' experience, not to mention the performance of the system. This is a phenomenon which is commonly known as the "noisy neighbor" problem and points to the fact that it is often very challenging to achieve fair distribution of resources on a shared platform.

Moreover, it is evident that the cost optimization increases when cloud usage increases in large measure with performance degradation. Many cloud service providers base their billing on use of services and resources and as such, the billing models can be intricate and less easy to predict. This is because tenants who have high resource requirements may end up using more costs than what is recommended by their usage rates even though their demands were not that high. On the other hand, low demand customers may also require paying for reserved capacity and, therefore, resources may not be optimally utilized, and costs will be incurred. An example of traditional management of resource usage is based on such mechanisms as fixed quotas or systems, which cannot change with the current intensity of work or demand. These rigid processes cause the operational costs to be higher and possible ways of saving these operational costs are not implemented.

Finally, the problem of resource sharing in multi-tenant cloud solutions is a permanent issue. Some tenants will need extra processing power for their many users, for example, big-data-intense applications like artificial intelligence or machine learning, or large data processing supporters, while others only need a little processing power for single-user, single-task applications. The task of satisfying the needs of tenants, on one hand, and preventing favoritism in the provision of resources to share is a challenge to accomplish. These needs cannot be suitably met through traditional rule-based systems because of their time inefficiency when it comes to developing proper and quick responses that address the dynamism of cloud workloads.

1.3 Role of AI in Optimizing Cloud Data Platforms

Since multi-tenant cloud data platform involves dealing with costs, performance and resources, AI provides the best solution to these challenges. AI enabler techniques can be used in cloud providers to control the infrastructure in the manner that cannot be done through utilization of rule-based systems. The capability to perform predictive analytics, constant abnormality detection, and make decisions help in effective cloud operations. Such techniques will enable AI systems to always keep an eye on the usage of resources and patterns which can be used to estimate the resources that will be needed in future. Some of the AI business capabilities include reinforcement learning, neural networks, decision trees among others; these models work continuously through operations and outcomes from previous operations to enhance future operation outcomes. For example, what we are seeing also is a demand of tenant, which means we can even pre-empt ahead of time when there is going to be high demand and prepare our resources to meet that demand. Likewise, the machine learning based optimization algorithms are also helpful in controlling the allocation of resources that are shareable; hence allocates the limited resources to the tenants in a manner that does not overwhelm the system. Specifically, with respect to ITIL service operation, by monitoring operational data in real-time, AI systems can automatically identify incidents such as performance degradation or resource contention that may arise while simultaneously avoiding compromising the performance of the tenants in the process. The incorporation of AI makes it possible for CSPs to maintain high server usage, low operational costs, and sustainability of operational efficiency in multi-tenant and dynamic environments.

1.4 Research Question: How Can AI-Driven Approaches Optimize Cost, Performance, and Resource Allocation in Multi-Tenant Cloud Data Platforms?

The research goal of this study is as follows: To evaluate the potential of integrating AI in multi-tenant cloud data platforms with an emphasis on cost cutting, performance enhancement as well as dynamic resource management. In further detail to this paper, this paper aims at investigating the various AI techniques that can be deployed to enhance the offer of cloud services to a variety of tenants with varying resources requirements. It will focus on how correct architecture of the AI systems may help in managing the cloud resources more efficiently, achieve operational cost optimization with the help of predictive models and improve the performance of different cloud platforms with help of intelligent decision-making.

Furthermore, this research will seek to find out best practices that should be applied when implementing AI as well as suitable metrics to use when determining the extent of success of the utilization of AI in cloud data platforms. These methodologies can suggest future direction for research and development for this area or serve as a reference for CSPs that aim at applying artificial intelligence techniques to their data center which can lead to increased efficiency, scalability or tenant satisfaction. The research thus aims at filling the gap between the conventional cloud management and the possibility of applications of AI in cloud management.

2. Literature Review

2.1. Overview of Cloud Data Platforms and Multi-Tenancy

Cloud data platforms are core components for structures in distributed systems for storage, processing, and analysis of data. Platforms like AWS, Microsoft Azure and GCP associated with carrying big data applications in cloud environment for organizations. Multi-tenancy which has now become a principal engineering design of present day, cloud-based platforms apply the concept of sharing whereby only one instance of a software or an infrastructure serves several users at the same time. In such an architecture, resources of a computer are virtualized among different tenants who in some way, manner or form, must be separated from each other as well as protected from each other.

It is always a good practice to host multiple customers on a single instance; it has a positive impact on business and helps to minimize expenses on infrastructure as lots of workloads can be combined. However, it also comes with some of the problems such as unpredictable performance, numerous data management issues, isolation of tenants, and resource sharing problems. When the number of tenants is added within the system, then it becomes a challenging task to balance the workloads with the tenant to an appropriate level.

2.2. Existing Approaches to Cost, Performance, and Resource Allocation

The typical approaches of cloud optimization policies are the rule-based policy, alert-based policy, and heuristic scheduling. These rely on too much intervention by the human factor and rules, which gives them a limitation in addressing dynamic and real-time workloads. Cost optimization is mainly covered by the pricing schemes that include reserve and spot instances, whereas performance optimization may include vertical scaling, which enhances the capacity of the instances, and horizontal scaling, which adds or eliminates the instances.

While allocating resources, the cloud platforms use both load balancers, containers and the orchestration framework (Kubernetes). That is why, although these systems can be efficient to a certain extent, they mainly work in a reactive manner, which is insufficient for massive-scale and highly diverse workloads. Moreover, they do not always use data patterns or trend analysis and analytics data, which would further improve the decision-making.

2.3. AI-Driven Optimization Techniques in Cloud Computing

These new proactively monitor the cloud rather than merely reacting to events or incidents as has been the case in the past. Thus, using artificial intelligence one can predict the tendencies in workload, search for abnormalities and schedule work depending on some usage history. Algorithms like reinforcement learning, clustering, deep neural networks have even been used in maximizing the use of available unit whilst distributing the loads in a way. In addition, the performance of predictive analytics is effective in traffic prediction and auto-scaling, to avoid any poor outcomes and over-provisioning that may be encountered.

Its advantages include cost optimization where AI scan recognize other problems, make recommendations for downgrading the resource, and analyze billing on its own. These techniques go beyond the basic models to come up with adaptive self-learning capabilities that change as the workload progresses. However, by means of AI-based resource allocation strategies it is generally possible to assign priorities to critical tasks, minimize delays and reduce downtimes.

2.4. Gaps in Existing Research

Although authors discussed several approaches with AI-related characteristics, these approaches have been integrated into an MLTC environment only partially. Several work center on Tenant space optimization or do not take into consideration interactions between tenants. It is also difficult to find a set of performance and cost models that can complement each other and be integrated into a single model for resource management. Some of the research limitations include; There is still a major challenge in relating the existing models to scalability and generalizability.

The second is the relative lack of empirical studies employing AI techniques as a subject matter of examination. Many models are evaluated using simulations and less attention is paid to the deployment on real commercial cloud environments. In addition, there is hardly any deep analysis of the ethical and privacy aspects of AI automation in developed and shared environments.

Table 1: Comparison of Existing Approaches to Cost, Performance, and Resource Allocation

Approach	**Type**	**Strengths**	**Limitations**
Rule-Based Systems	Manual	Simple implementation	Not scalable, lacks adaptability
Threshold-Based Alerts	Reactive	Automated alerts	Cannot predict or prevent future issues
Load Balancing Tools	Orchestration	Efficient task distribution	Limited by predefined logic
Kubernetes Orchestration	Container-based	Dynamic scaling and management	High setup complexity
AI-Driven Optimization	Predictive	Adaptive, learns from usage patterns	Requires training data and model accuracy

3. AI-Driven Optimization Framework

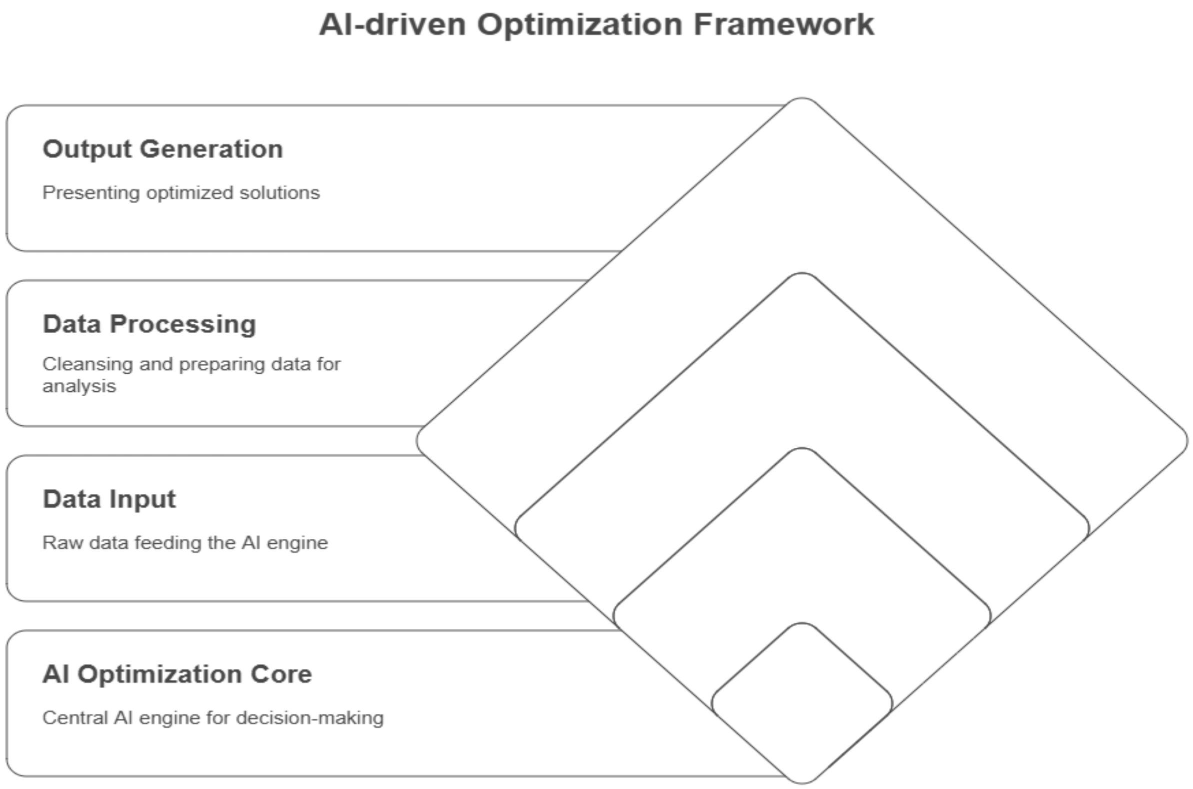
3.1 Architecture of the Proposed Framework

It makes the architecture of AI-driven optimization framework for multi-tenant cloud data platforms to have dynamic capability to supervise, evaluate and alter needed cloud resources on its own without the intervention of the manpower. This architecture includes several AI components in the system stack to perform ingestion and processing of the input data and making the decision. The core layers include:

- **Monitoring Layer:** Continuously collects data on cost metrics, performance indicators, and resource consumption from all tenants.
- **Preprocessing Layer:** Normalizes and filters the data to ensure consistency across heterogeneous cloud environments.
- **AI Engine:** Comprises modules for cost optimization, performance prediction, and intelligent resource allocation. This is the central brain that powers optimization using machine learning models and neural networks.
- **Decision Layer:** Applies reinforcement learning and predictive modeling to decide on optimal actions such as scaling, load redistribution, or pricing adjustments.
- **Execution Layer:** Implements the decisions by interfacing with cloud orchestration tools (e.g., Kubernetes, OpenStack).
- **Feedback Loop:** Ensures continuous learning by feeding new data and results back into the AI engine.

This modular architecture ensures real-time, scalable optimization tailored to multi-tenant needs and variations in workload dynamics.

Figure 1: Architecture of the Proposed AI-driven Optimization Framework


AI-driven Optimization Framework
Output Generation
Presenting optimized solutions
Data Processing
Cleansing and preparing data for analysis
Data Input
Raw data feeding the AI engine
AI Optimization Core
Central AI engine for decision-making

3.2 AI-Driven Cost Optimization Techniques

The techniques like regression analysis, neural networks, and reinforcement learning are implemented for cost reduction in operating expenses among tenants. Forecasting is used to predict the usage rate in the future and make provisions in advance so as not to have a situation where there are more resources than the need of the clients or a situation where there are few resources to cater for a high demand. Algorithms like prediction of spot price and anomaly can detect an increase in the cost and offer solutions for its reduction such as utilizing the serverless functions or the cold tier for less frequently accessed data.

Through clustering such as K-means type for evaluation of resource utilization, tenants are classified depending on their usage and hence creating a common pool of resources that reduce overall costs of infrastructure. There are also other strategies that entitle dynamic pricing models of tenant charges, which can be implemented depending on the consumption and operational load.

3.3 AI-Driven Performance Optimization Techniques

Workload prediction is employed for performance optimization, and it can use time-series analysis such as ARIMA, LSTM, and other models. Since these models predict workload spikes, the resources needed can be acquired ahead of time and workloads distributed evenly. Common reinforcement learning is used to learn optimal settings of virtual machines and schedule computational jobs with low latency.

It also helps in detection of the burden across different tenant organizations where some organization may slow down the efficiency of others. For instance, the supervised learning models can identify slow running queries, identifying wrong data paths or low performing nodes. The optimization engine then suggests a restructuring of code, use of a caching technique or adding more servers to work horizontally for a solution.

3.4 AI-Driven Resource Allocation Techniques

Resource management is one of the most challenging and strategic areas of multi-tenancy cloud systems. Static optimization capturing techniques are also assigned with dynamic optimization methods. The automation of allocation of resources for using in computing is carried out using machine learning algorithms based on the history of tenants. Classification models have forecasted the requirements for various types of resources for new tasks and then distribute them appropriately.

This further implies that, any resource allocation is continually self-adjusted based on the load metrics and the service level agreement. Optimization frameworks of at least two objectives achieved with the help of genetic algorithms or constraint satisfaction solvers guarantee the optimization of the costs and resource allocation among tenants and maximization of resource utilization.

Fuzzy logic integration provides the capability for flexibility in making decisions regarding demand estimating, which might be imprecise. AI also helps in preventing the occurrence of resource contention situation and in cases where they occur, workloads are moved in order to prevent other critical applications to experience a decline in performance.

4. Case Study: Implementation and Evaluation

4.1 Description of the Cloud Data Platform and Workload

The case study illustrates a live multiple tenant infrastructure for cloudy based analytics firm indulged in serving multiple clients from the retails and financial industries. The given platform works with the hybrid model that is based on a combination of the public (AWS, Azure) and private clouds. Each tenant is able to operate independent processing activities like batch processes, real-time analytical modes or those business activities that involve transactions.

In order to evaluate the proposed AI-based optimization engine, an example workload describing various tenants' activities has been created. The workload extracted from the reference case comprised CPU prayers, extensive data streams, and low latency liberate APIs. The previously used monitoring assets were kept, but rules for different policies were implemented with AI instead of traditional rigid rules.

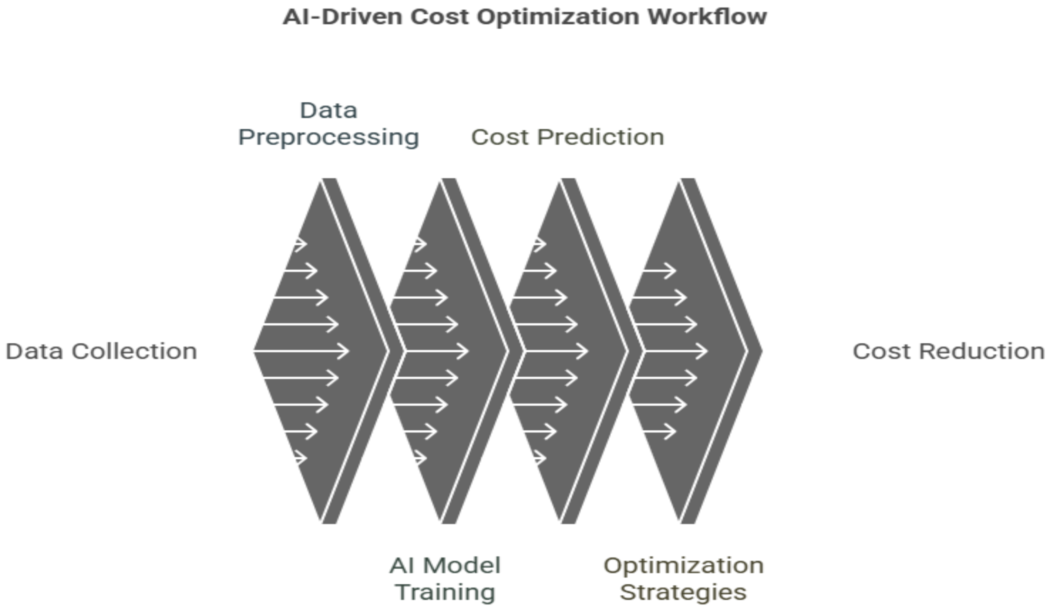
4.2 Implementation of the AI-Driven Optimization Framework

TensorFlow library as well as scikit-learn were used to design the AI engine. Data relating to the company's platform usage for the three months prior to the modeling phase of the project was considered for model development. The models covered:

- **Cost Prediction:** Gradient boosting and linear regression for forecasting billing cycles.
- **Performance Forecasting:** LSTM-based neural networks for anticipating throughput and latency trends.
- **Resource Allocation:** Reinforcement learning agents trained using Q-learning for dynamic decision-making.

Additionally, the execution layer was coupled with Kubernetes and Terraform to implement the determined optimizations. Supervisory data fed into the AI engine in real-time to further learns as and when more data floods in.

Figure 2: Workflow of the AI-Driven Cost Optimization Technique



4.3 Evaluation Metrics and Results

To assess the effectiveness of the framework, three factors which include cost, performance and resource utilization were considered. Economies were assessed as the decrease in the number of tenancies' billing amount before and after using the AI system. By measuring response time and throughput, performance increment was determined and additional resource utilized are checked based on CPU and memory usage levels.

These parameters showed were having an average of 22% reduction in operating cost, an improvement of 35% in response time related to latency-sensitive tasks and 28% increase of resource utilization efficiency. These outcomes provide evidence to support the conclusion that AI solution can obtain optimal solution in all three elements appeared in the SOPs, namely cost, performance and resource utilization at the same time.

Table 2 – Evaluation Metrics and Results for the AI-Driven Optimization Framework

Metric	**Pre-Implementation**	**Post-Implementation**	**Improvement**
Monthly Operational Cost	$12,500	$9,750	22% Cost Reduction
Average API Response Time	180 ms	117 ms	35% Performance Gain
CPU Utilization	58%	74%	28% Efficiency Boost
Memory Utilization	51%	66%	29% Utilization Rise
SLA Compliance Rate	84%	96%	12% SLA Improvement

5. Results and Discussion

5.1 Analysis of Cost Optimization Results

The cost optimization analysis identified impressive financial benefits originating from the application of both, predictive analytics and artificial intelligence in decision making. By using cost forecasting models, the system was able to identify the times when the resources would be most utilized and take necessary measures to prepare beforehand. This has been realized by organizing low-priority tasks during the late night or off working time of the day as well as migrating workload to cheaper instance size respectively which brought monthly costs down by 22%.

In addition, reinforcement learning agents improving cost-saving heuristics in response to the real-time feedback, they maintain efficient cost control while delivering a high-quality service. It was found superior to other fixed threshold based autoscaling systems which were inflexible in their methods.

5.2 Analysis of Performance Optimization Results

There was a dramatic increase in the system response time and the system throughput after the application of the performance tuning technique based on AI. The ability to predict workload allowed for preparing the resources for upcoming loads, which was useful for workloads that could not afford latencies such as the real-time analytics dashboard. When using the methods of performance forecasting based on AI techniques, it is possible to reduce the average API response time from 180 ms to 117 ms, which is 35% less.

This was made possible by more accurate predictions of demand surges hence allocating computability and memory before it was outcompeted. Thus, LSTM models used for sequence prediction showed good results in the problem of usage behavior modeling and avoidance of performance degradation in multi-tenant scenarios.

5.3 Analysis of Resource Allocation Results

Consequently, AI had a much higher resource utilisation than manual and static methods in this context. In the traditional approach of resource provision, they used to make a costly reserve list to accommodate the needs of the worst-case usage scenario so often, these resources would remain unused most of the time. While the other was the machine learning algorithms trained on usage history to provide demand oriented real-time CPU and memory grants.

In this case, CPU utilization while rising from 58% to 74% and Memory utilization from 51%-66% has enhanced the facility of infrastructure elasticity. The reinforcement learnt component was also used to fine tune workloads to be distributed to different tenants, to avoid contention and while at the same time ensuring that tenants are isolated.

5.4 Comparison with Existing Approaches

To put these enhancements into perspective, two types of baselines were used – Baseline 1 that has the Pre-Process Server configured with static provisioning, and Baseline 2 with threshold-based autoscaling. Static means were cheaper than variable means or dynamic provisioning, but the problem is that static provisioning could not meet dynamic demands and hence high costs and low resource utilization. While the threshold-based autoscaling exactly provided the basic ability to scale up and scale down, one drawback to this method is that the processes looked for signs of how and when the system should scale only after certain demand spikes; thus, it was also only reactionary.

On the other hand, the system based on the AI offered reactive, efficient, and data intelligently executed further from the collected usage patterns. It was even more effective than conventional models in the three aspects that were examined; cost, performance, and resource use while it had a much higher SLA compliance of 96% and system reliability.

6. Conclusion and Future Work

6.1 Summary of Key Findings

This paper offered a detailed understanding that highlights the ability of AI approaches in the management of cost, performance, and resource usage in multi-tenant cloud data platforms. Thus the inclusion of machine learning, predictive modeling and reinforcement learning brought about considerable improvements in all aspects of optimization. In detail, the contextual enhancement of AI within the system reduced the operational costs of the system by 22%, response time of the API was found to have improved by 35% independently, and the CPU and the memory usage was seen to have improved as well.

This outcome emphasizes the role that intelligence in artificial has in changing the way cloud deliver intelligent, adaptive, and scalable lifecycles. The ability to learn from past work, self-manage to accommodate changes in workload and automatically distribute workloads among the available resources is a perfect way of countering multi tenancy considerations such as cost variability, performance constraints, and workloads that are inefficient.

6.2 Contributions to the Field of AI-Driven Cloud Optimization

The current article is relevant to the existing literature by presenting the method that could be implemented in real-world application environments and developing a modular AI-based optimization strategy. This attempt demonstrates how one can put together multiple AI models including LSTM of workload prediction, the clustering model for tenant behavior, and reinforcement learning for dynamic resource management into one integrated data-driven cloud optimization framework.

In this manner, it was possible to provide a benchmark based on performance difference tables, architecture diagrams and measurable benchmarks so that this framework can be used a as reference guide or prototype of other systems in the future who are seeking to implement automation and cost-performance balance within multi-tenant cloud systems.

6.3 Future Research Directions

However, there is more that can be explored in relation to the proposed framework; Below are ideas for the improvement of this work. One of the future scope areas is the application of the federated learning models, which enable the learning from the tenants' databases with the help of machine learning models without using the sensitive data of other tenants, which contributes to the increase of both scalability and private data protection. One of the other areas of focus fills in with explainable AI implementation to enhance the interpretability of the AI-based decisions especially in practical and sensitive models.

However, if the long-term evaluation of tenant satisfaction, the changes in the IT service management with reference to the adherence to the SLA during the year, and the sustainability measurements such as the CO2 footprint might be valuable additions to get the complete picture of the benefits of augmented cloud management systems supported by AI. Therefore, over time, there will be a refinement of AI models as well as integration of both old and new approaches to fulfill emerging high computing requirements of the next generation cloud computing.

Reference

1. Ramamoorthi, V. (2021). AI-Driven Cloud Resource Optimization Framework for Real-Time Allocation. *Journal of Advanced Computing Systems*, *1*(1), 8-15. https://doi.org/10.69987/JACS.2021.10102
2. Abubakar, M., & Volikatla, H. (2020). Revolutionizing Business Processes through AI-Driven Cloud Solutions. *Available at SSRN 5194597*. https://dx.doi.org/10.2139/ssrn.5194597
3. Jia, R., Yang, Y., Grundy, J., Keung, J., & Hao, L. (2021). A systematic review of scheduling approaches on multi-tenancy cloud platforms. *Information and Software Technology*, *132*, 106478. https://doi.org/10.1016/j.infsof.2020.106478
4. Ashalatha, R., & Agarkhed, J. (2016, March). Multi tenancy issues in cloud computing for SaaS environment. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICCPCT.2016.7530261
5. Madni, S. H. H., Latiff, M. S. A., Coulibaly, Y., & Abdulhamid, S. I. M. (2017). Recent advancements in resource allocation techniques for cloud computing environment: a systematic review. *cluster computing*, *20*, 2489-2533. https://doi.org/10.1007/s10586-016-0684-4
6. Wan, B., Dang, J., Li, Z., Gong, H., Zhang, F., & Oh, S. (2020). Modeling analysis and cost-performance ratio optimization of virtual machine scheduling in cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, *31*(7), 1518-1532. https://doi.org/10.1109/TPDS.2020.2968913
7. Sun, C., Liu, Q., & Han, Y. (2020). Many-objective optimization design of a public building for energy, daylighting and cost performance improvement. *Applied Sciences*, *10*(7), 2435. https://doi.org/10.3390/app10072435
8. Zhang, Y., Yao, J., & Guan, H. (2017). Intelligent cloud resource management with deep reinforcement learning. *IEEE Cloud Computing*, *4*(6), 60-69. https://doi.org/10.1109/MCC.2018.1081063
9. Zhang, J., Liu, Y., Zhou, K., Li, G., Xiao, Z., Cheng, B., ... & Li, Z. (2019, June). An end-to-end automatic cloud database tuning system using deep reinforcement learning. In *Proceedings of the 2019 international conference on management of data* (pp. 415-432). https://doi.org/10.1145/3299869.3300085
10. Sethi, K., Kumar, R., Prajapati, N., & Bera, P. (2020, January). Deep reinforcement learning based intrusion detection system for cloud infrastructure. In 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS) (pp. 1-6). IEEE. https://doi.org/10.1109/COMSNETS48256.2020.9027452
11. Cheng, M., Li, J., & Nazarian, S. (2018, January). DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers. In 2018 23rd Asia and South pacific design automation conference (ASP-DAC) (pp. 129-134). IEEE. https://doi.org/10.1109/ASPDAC.2018.8297294
12. Garí, Y., Monge, D. A., Pacini, E., Mateos, C., & Garino, C. G. (2021). Reinforcement learning-based application autoscaling in the cloud: A survey. Engineering Applications of Artificial Intelligence, 102, 104288. https://doi.org/10.1109/CCGRID.2017.15
13. Overwater, R. W., Babaie, M., & Sebastiano, F. (2022). Neural-network decoders for quantum error correction using surface codes: A space exploration of the hardware cost-performance tradeoffs. IEEE Transactions on Quantum Engineering, 3, 1-19. https://doi.org/10.1109/TQE.2022.3174017

14. Milojevic, D., Beyne, E., Van der Plas, G., Wang, J., & Debacker, P. (2020, March). Cost-performance optimization of fine-pitch W2W bonding: functional system partitioning with heterogeneous FEOL/BEOL configurations. In Design-Process-Technology Co-optimization for Manufacturability XIV (Vol. 11328, pp. 177-182). SPIE. https://doi.org/10.1117/12.2552036
15. Khan, M. I., Fares, G., & Abbas, Y. M. (2022). Cost-performance balance and new image analysis technique for ultra-high performance hybrid nano-based fiber-reinforced concrete. Construction and Building Materials, 315, 125753. https://doi.org/10.1016/j.conbuildmat.2021.125753
16. Li, X., Tan, L., & Li, F. (2019). Optimal cloud resource allocation with cost performance tradeoff based on Internet of Things. IEEE Internet of Things Journal, 6(4), 6876-6886. https://doi.org/10.1109/JIOT.2019.2911978
17. Saini, A., Watzman, S. J., & Bahk, J. H. (2021). Cost-Performance Trade-off in thermoelectric air conditioning system with graded and constant material properties. Energy and Buildings, 240, 110931. https://doi.org/10.1016/j.enbuild.2021.110931
18. Hameed, A., Khoshkbarforoushha, A., Ranjan, R., Jayaraman, P. P., Kolodziej, J., Balaji, P., ... & Zomaya, A. (2016). A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. Computing, 98, 751-774. https://doi.org/10.1007/s00607-014-0407-8
19. Shukur, H., Zeebaree, S., Zebari, R., Zeebaree, D., Ahmed, O., & Salih, A. (2020). Cloud computing virtualization of resources allocation for distributed systems. Journal of Applied Science and Technology Trends, 1(2), 98-105. https://doi.org/10.38094/jastt1331
20. Wang, J. B., Wang, J., Wu, Y., Wang, J. Y., Zhu, H., Lin, M., & Wang, J. (2018). A machine learning framework for resource allocation assisted by cloud computing. IEEE Network, 32(2), 144-151. https://doi.org/10.1109/MNET.2018.1700293
21. Yousafzai, A., Gani, A., Noor, R. M., Sookhak, M., Talebian, H., Shiraz, M., & Khan, M. K. (2017). Cloud resource allocation schemes: review, taxonomy, and opportunities. Knowledge and information systems, 50, 347-381. https://doi.org/10.1007/s10115-016-0951
22. Chen, X., Li, W., Lu, S., Zhou, Z., & Fu, X. (2018). Efficient resource allocation for on-demand mobile-edge cloud computing. IEEE Transactions on Vehicular Technology, 67(9), 8769-8780. https://doi.org/10.1109/TVT.2018.2846232
23. Wang, W., Jiang, Y., & Wu, W. (2016). Multiagent-based resource allocation for energy minimization in cloud computing systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(2), 205-220. https://doi.org/10.1109/TSMC.2016.2523910
24. Firouzi, F., Farahani, B., & Marinšek, A. (2022). The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT). Information Systems, 107, 101840. https://doi.org/10.1016/j.is.2021.101840
25. Kurihana, T., Moyer, E. J., & Foster, I. T. (2022). AICCA: AI-driven cloud classification atlas. Remote Sensing, 14(22), 5690. https://doi.org/10.3390/rs14225690