World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Dynamic resource allocation using AI-driven workload forecasting in multi-cloud environments

Adetayo Adeyinka *

Independent Researcher, USA.

World Journal of Advanced Research and Reviews, 2024, 23(01), 3188-3198

Publication history: Received on 08 June 2024; revised on 22 July 2024; accepted on 28 July 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.1.2178

Abstract

This research investigates the application of artificial intelligence (AI) for dynamic resource allocation using workload forecasting in multi-cloud environments. With the growing adoption of multi-cloud strategies, organizations face increasing challenges in managing resource distribution efficiently due to fluctuating and unpredictable workloads. To address this, the study introduces an AI-driven framework that combines time-series forecasting models such as Long Short-Term Memory (LSTM) networks, reinforcement learning, and decision tree-based algorithms to accurately predict workload demands and allocate resources dynamically across multiple cloud platforms. The system continuously monitors workload patterns and adjusts resource provisioning in real-time to enhance performance and cost-efficiency. Experimental results demonstrate that the proposed approach significantly improves CPU and memory utilization, reduces operational costs by up to 25%, and increases SLA compliance. By offering a scalable, intelligent solution for resource management, this research contributes to the advancement of autonomous cloud operations. It provides practical value for optimizing complex multi-cloud infrastructures' performance, reliability, and efficiency.

Keywords: Cloud Computing; Multi-Cloud Environments; Resource Allocation; Workload Forecasting; Artificial Intelligence (Ai)

1. Introduction

1.1. Background on cloud computing and resource management

Cloud computing is a new IT technology known as the third revolution after personal computers and the internet. The enhancement and development of distributed databases, parallel computing, grid computing, and distributed computing resulted in cloud computing. In the 1960s, John McCarthy envisioned that computing facilities would be provided to the general public, like a utility. Cloud computing is a platform based on the internet that offers myriads of services based on plug-and-play.

* Corresponding author: Adetayo Adeyinka

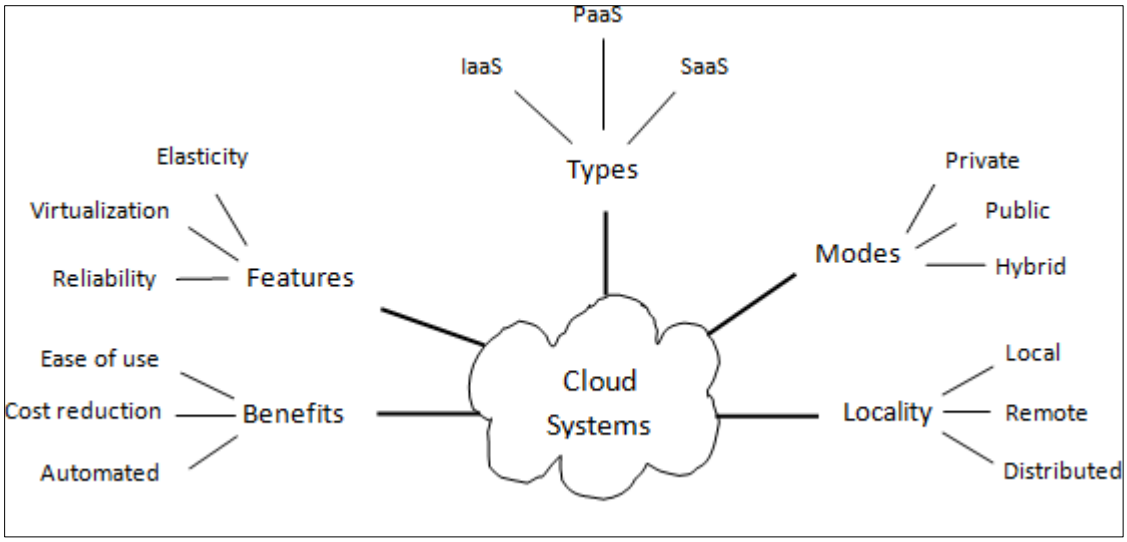

Figure 1 Cloud Computing System

National Institute of Standards and Technology refers to the inclusion of common key elements that are massively used in cloud computing. "It is a model for enabling convenient, on-demand networking access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction." It refers to some key elements which need to be analyzed first. Typically, cloud computing provides several benefits that could not otherwise be realized, including scalability and service quality. Cost-effective cloud computing offers a specialized environment and a simplified interface. When concerned about cloud computing, we understand that resource management allocates computing storage, networking, and resources to a set of applications, cloud service providers, and cloud users. The cloud computing paradigm has emerged, wherein a pool of computing resources is shared between the applications that may be accessed over the internet. The technology community and the common public use this term. Our objective with the paper is to conduct an inclusive survey of recent research into the challenging resource management genre in a cloud environment. Conveying upon the complexity of the problem, describe the state of the art and outline the fundamental open challenges.

1.2. Challenges in resource allocation in multi-cloud settings

One of the major challenges in cloud computing is the allocation of resources due to their dynamic nature. End-users can access resources ubiquitously—at any time and from anywhere—thanks to the scalability of cloud services. Although cloud data centers provide a large pool of computing resources and support on-demand, flexible allocation, the abundance of these resources can sometimes lead to inefficient or non-optimal allocation. The core objective of resource allocation for cloud brokers may involve optimizing application performance, improving energy efficiency, or maximizing resource utilization. This applies to all forms of ICT resources, such as elastic IPs, blob storage, and servers, which are assigned to end-users.

Energy-efficient resource allocation introduces several challenges. These include selecting the most suitable workload and interference categories based on power consumption, performance, and resource usage. Improving the cloud infrastructure layout—through better technologies, tools, topology design, and resource management—is also crucial. Another important aspect is resource consolidation, which involves analyzing application interdependencies to enhance performance and return on investment. Furthermore, evaluating standardized, federated, and centralized data center resources is essential for efficient provisioning and runtime resource utilization.

Additional concerns include ensuring system resilience and failure recovery by improving network availability, asset utilization, and power efficiency and reducing the time required to allocate resources. Another priority is designing practical, elastic cloud infrastructure supporting business continuity and mission-critical workloads. In this context, a "resource" in cloud computing refers to any core infrastructure component, including storage, CPU, bandwidth, and memory.

Resource allocation is a highly complex problem. It often requires assumptions such as the cost of utilized resources, minimizing energy waste, and identifying active server sets. Cloud brokers are responsible for assigning system resources to CPUs and must also determine whether to accept incoming service requests based on the availability of

resources. Several interdependent factors complicate this task—these include monitoring resource availability, tracking service requirements, analyzing user demand, calculating resource costs, managing performance levels, and observing the actual distribution of resources. These elements contribute to the multifaceted nature of resource allocation in cloud computing environments.

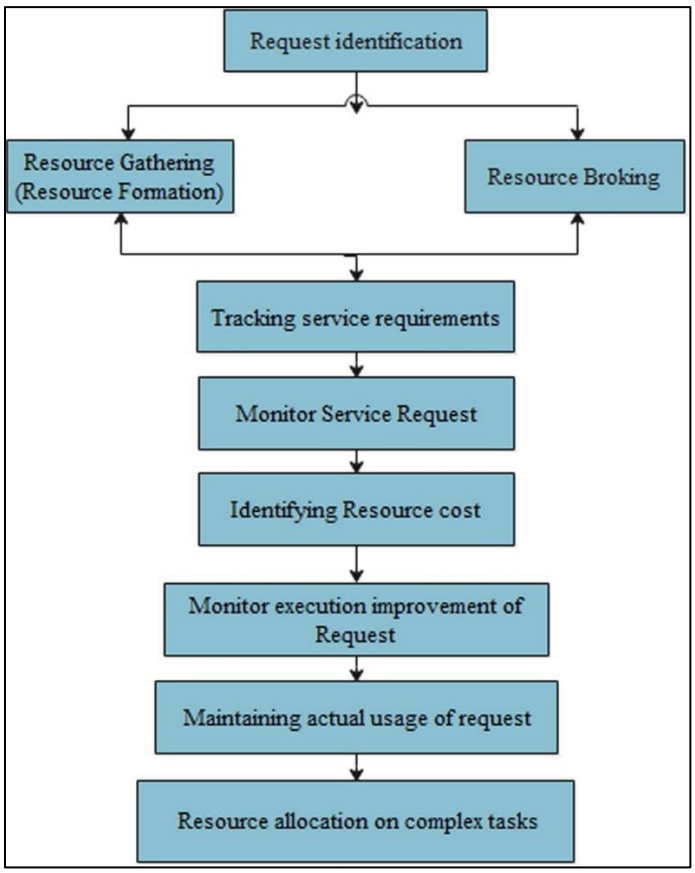

Figure 2 Challenges of resource allocation assignment in the cloud

2. Importance of workload forecasting in dynamic environments

A crucial component of successfully managing cloud infrastructures is workload forecasting. It entails anticipating the demand for computing resources to manage resources and guarantee optimal utilization appropriately. This review gives a thorough grasp of workload forecasting methods and the significance of those methods in cloud computing settings. In recent years, we have seen a meteoric rise in cloud usage, resulting in an exponential increase in the number of individuals and applications served on cloud platforms. To avoid the inadequacy or over-provision of capital in response to this rise in demand, precise workload forecasting is necessary. Incorporating advanced algorithms that modify and automatically adjust depending on actual and historical information is another effort to increase prediction accuracy. Cloud service providers may efficiently distribute resources, guarantee excellent system performance, boost cost-effectiveness, and play a crucial role in workload forecasting for cloud-based systems by properly anticipating workload trends. [32] The expected workload of web platform services on cloud computing is essential to maintaining service effectiveness. It is important to operate web platform services as high pressure on the service can slow down dedicated servers. Speed and support decline as more customers try to access the web platform's service. This method reflects the importance of great forecasting and regulation for cloud web platforms. [33]

2.1. Role of Artificial Intelligence (AI) in predictive analytics and automation

In today's data-driven world, businesses increasingly rely on predictive analytics and forecasting to maintain a competitive edge, and Artificial Intelligence (AI) plays a central role in this shift. AI enables organizations to process vast volumes of data, recognize patterns, and generate more accurate predictions, significantly transforming decision-making processes. As a leading data insights partner, Gate6 leverages AI to enhance predictive analytics and forecasting capabilities, delivering actionable insights that support smarter business strategies. One of the key advantages of AI is

its ability to process and analyze large and complex datasets far more efficiently than traditional methods. AI algorithms can quickly sift through structured and unstructured data, uncovering trends and relationships that human analysts may overlook. This saves valuable time and ensures that business decisions are based on comprehensive and up-to-date information.

Additionally, AI's capability for real-time analysis allows companies to respond swiftly to dynamic market conditions. By continuously monitoring data as it flows in, AI-powered systems provide immediate insights, enabling timely adjustments such as inventory changes or reactions to emerging trends. Furthermore, AI-driven automation streamlines repetitive tasks like data collection, cleaning, and analysis. This reduces the risk of human error and frees up resources for more strategic initiatives while improving accuracy and scalability in analytics. Altogether, AI transforms predictive analytics into a faster, more precise, and more scalable tool for business success.

2.2. Problem Statement

In today's rapidly evolving digital landscape, cloud computing is critical in delivering scalable and flexible infrastructure services. However, one of the most pressing challenges in multi-cloud environments is the dynamic and efficient allocation of computational resources. Traditional resource allocation strategies, often static or reactive, cannot effectively respond to unpredictable workload variations, leading to underutilization, performance degradation, increased operational costs, and service-level agreement (SLA) violations. These limitations are exacerbated in multi-cloud settings where resource heterogeneity and interoperability add further complexity. Therefore, there is a growing need for intelligent, proactive systems capable of forecasting workload demands and autonomously adjusting resource allocation strategies in real time.

2.3. Objectives of the Study

This study addresses the challenges above by exploring how artificial intelligence (AI), particularly time-series forecasting and reinforcement learning, can be integrated into resource allocation mechanisms to enhance efficiency, scalability, and reliability in multi-cloud environments. The specific objectives of the study are:

- To develop a predictive model using AI techniques (e.g., LSTM, ARIMA, Prophet) to forecast cloud workload patterns accurately.
- To implement a dynamic resource allocation framework based on reinforcement learning and intelligent scheduling algorithms.
- To evaluate the proposed system's performance regarding cost-efficiency, resource utilization, latency, and forecasting accuracy.
- To compare the effectiveness of AI-driven methods against traditional resource allocation baselines within simulated and real-world multi-cloud infrastructures.

2.4. Structure of the Paper

2.4.1. The paper is organized as follows

- Section 1 provides the background and context of dynamic resource allocation in cloud computing, including the motivation and significance of AI-driven approaches.
- Section 2 comprehensively reviews related literature covering traditional and AI-based resource management strategies.
- Section 3 details the research methodology, including the design of the AI models, workload simulation, and deployment setup across multi-cloud environments.
- Section 4 outlines the data used, the forecasting and allocation models implemented, and the evaluation metrics employed.
- Section 5 discusses the experimental results, comparing AI-based methods with conventional techniques.
- Section 6 interprets the findings, highlights the contributions, and identifies limitations and implications.
- Section 7 concludes the paper with a summary of key insights and recommendations for future research.

3. Literature Review

3.1. Review of traditional resource allocation techniques in cloud environments

As multi-cloud environments keep evolving, resource allocation has become crucial for utilizing application performance and the cost-effectiveness of cloud resources for running applications. Traditional resource allocation

methods follow a fix-the-best service approach, where multiple resources are required to access a single resource. However, within multi-cloud environments, the above approach is ineffective and inefficient. Public and multi-cloud service providers are constantly expanding their service offerings and minimizing costs by innovating and coordinating various groups within the companies.

These strategies and offerings change frequently within the industry and require a deep understanding of a cloud computing environment, its organizational and operational procedures, and the various interdependencies. Traditional resource allocation methods and approaches neither consider dynamic managerial needs nor different performance objectives. These, in turn, result in allocating and reallocating resources to meet organizational needs and goals over time. Existing standalone resource allocation models are impractical and unrealistic because they are performed in isolation from the tools, technologies, expertise, and partners, and they are also driven by variables external to the service system.

Thus, a dynamic allocation model based on the external environment is proposed to consider dynamic managers' needs, such as the minimum level of resources available to secure the continuation of services until the first load is induced into the system and before other resources are reallocated. Dynamic resource allocation is the process of combining these service offerings and categories into one allocation process. The rationale for proposing dynamic resource allocation is to consider all relevant offerings simultaneously to yield a more realistic proposition of cloud services.

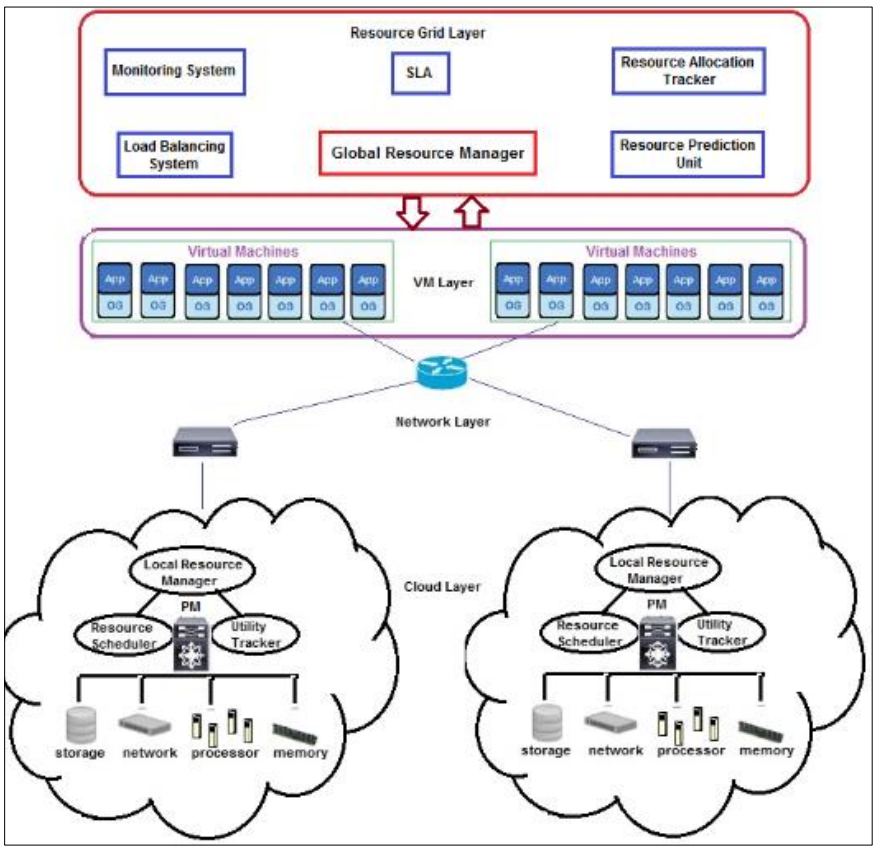

Figure 3 Multi-Cloud Environment

3.2. Studies on dynamic resource management in multi-cloud systems

Now, cloud computing is a business. So, the price and quality of service are important in this area. Cloudoviders provides resources and software to clients through vice-level agreements (SLA). SLA is a contract between the user and the cloud provider that contains the resource requirements of users, service time, and cost constraints of service, which have very important benefits for a business investor. Cloud service providers are willing to benefit from cloud computing service users and do not want to pay much. Cloud computing users receive good quality services from service providers, and the service fee is based on allocating resources in a particular service environment.

Cloud service providers must allocate resources to clients specific to a particular method [8]. There are several resource allocation models used in the field of cloud computing. Each model uses a specific technique and algorithms to achieve this goal. Consumers in cloud environments are not involved in significant investment in information technology (IT) infrastructures and complex issues related to their construction and maintenance. In this model, users can access needed services regardless of where services are hosted. Based on the "pay-use" model, cloud computing hosts practical, commercial, and scientific programs. Data centers hosting for applications consume a lot of energy. Maintenance of large data centers requires high energy consumption. It has been estimated that the cost of maintaining data centers increased the main cost of the original investment; in this case, the maintenance of data centers in this situation could be even adverse, and unfortunately, it has been seen that this process is being continued without any limit.

From the perspective of service providers, maximizing profits, given the high energy cost, is a problem. Since the profit is pared to expenses, its profit should be compensated by providing customer services. One way to reduce costs and increase profits is by reducing energy consumption. The rising energy cost is a big potential threat to increasing the cost of ownership.

3.3. Applications of AI and machine learning in cloud computing

Artificial intelligence and cloud computing are transforming industries through real-world applications across healthcare, finance, retail, and manufacturing. In healthcare, AI and machine learning are used for disease diagnosis and prognosis by analyzing medical images and patient data to detect conditions such as cancer, diabetes, and cardiovascular disease. For instance, Google Cloud's AutoML Vision enhances machine learning methods to identify issues in clinical images. Additionally, AI enables customized medical care by examining genetic information and patient records to develop personalized treatment plans, as seen in IBM Watson Health's use of algorithms to tailor therapies based on individual genetic profiles. In finance, AI supports fraud detection by analyzing real-time transaction patterns to identify suspicious behavior—Mastercard, for example, uses cloud-based AI to prevent fraudulent activities and save billions annually. Algorithmic trading is another application where firms like Alpaca use AI to analyze market trends and execute high-frequency trades at optimal times. In the retail sector, AI enables customer personalization by studying buying behavior to make tailored recommendations, with Amazon using cloud-powered systems to drive modern marketing and sales. AI also supports inventory management through demand forecasting, allowing retailers like Walmart to optimize stock levels, reduce costs, and avoid product shortages. In manufacturing, predictive maintenance powered by AI forecasts equipment failures before they occur, minimizing downtime and repair expenses—Siemens, for example, monitors equipment conditions through cloud data. Furthermore, quality control is enhanced as machine learning algorithms analyze production data in real time to detect defects; GE uses cloud-based tools to improve product quality and ensure compliance with manufacturing standards. These applications illustrate how AI and cloud technology drive efficiency, innovation, and accuracy across multiple sectors.

4. Methodology

4.1. Research Design

This study adopts a hybrid research design that integrates experimental, simulation-based, and real-time testing approaches to validate the effectiveness of AI-driven workload forecasting for dynamic resource allocation in multi-cloud environments. The experimental phase involves developing and training predictive models using historical cloud workload data. Simulation-based testing uses cloud infrastructure emulators to model different allocation strategies under controlled scenarios. Finally, real-time testing is implemented by deploying the models in a live multi-cloud environment to evaluate system responsiveness and adaptability under actual workload fluctuations. This layered approach comprehensively evaluates the proposed framework across synthetic and production conditions.

4.2. Data Collection

The research utilizes publicly available and proprietary datasets containing real-world cloud operational data. This includes cloud logs, CPU and memory usage patterns, application-level performance metrics, and workload traces collected over time from various cloud services. The datasets are sourced from established benchmarks such as the Google Cluster Data, Azure VM traces, and datasets from the Alibaba Cloud cluster. Data preprocessing involves cleaning, normalization, and feature extraction to make the data suitable for AI model training. The final dataset includes time-stamped metrics such as CPU utilization (%), memory consumption (GB), storage I/O rates, and bandwidth usage (Mbps), which are critical for forecasting workload trends and assessing resource demand.

4.3. AI Techniques Used

A suite of advanced artificial intelligence techniques is deployed to build a robust workload forecasting and resource allocation framework

- Time Series Forecasting: To predict future workload trends, the study employs Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) networks, and Facebook's Prophet model. ARIMA is used for linear pattern forecasting, while LSTM, a type of recurrent neural network, is applied to capture complex temporal dependencies in workload patterns. Prophet is incorporated for its scalability and flexibility in handling seasonality and trend shifts in large datasets.
- Reinforcement Learning (RL): Reinforcement learning agents are trained to learn optimal provisioning strategies based on reward feedback for dynamic and adaptive resource allocation. RL agents continuously interact with the simulated cloud environment, learning to allocate resources efficiently under varying load conditions while minimizing cost and SLA violations.
- Neural Networks & Hybrid Models: Feedforward neural networks are used for workload classification and short-term demand estimation. Hybrid models that combine forecasting and decision-making capabilities—such as integrating LSTM with Deep Q-Networks (DQN)—are also developed to improve system responsiveness and accuracy.

4.4. Workload Modeling

Workload modeling is central to understanding and simulating cloud behavior. Key performance metrics used to model workloads include

- CPU utilization (%)
- Memory usage (GB)
- Storage I/O (operations per second)
- Network bandwidth (Mbps)

These metrics are extracted and profiled to build workload signatures that reflect application demand characteristics. Workloads are categorized into compute-intensive, memory-intensive, and I/O-intensive classes to guide tailored resource provisioning.

4.5. Resource Allocation Algorithm

The dynamic resource allocation algorithm integrates both scheduling and load-balancing techniques

- Scheduling Models: Tasks are scheduled using heuristics such as First Fit Decreasing (FFD), Earliest Deadline First (EDF), and AI-augmented scheduling using predicted workload data. These algorithms prioritize task execution to maximize resource efficiency and minimize delay.
- Load Balancing and Scaling Policies: Load is balanced across multiple cloud instances using dynamic policies that respond to workload forecasts. Auto-scaling policies are defined based on forecasted CPU and memory thresholds, and container orchestration tools such as Kubernetes are used to manage horizontal and vertical scaling.

4.6. Simulation and Deployment Environment

The proposed solution is validated using a simulated multi-cloud environment comprising leading cloud platforms, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). CloudSim and iFogSim are used for simulation experiments, allowing the configuration of cloud nodes, VM instances, and network conditions. For real-time deployment, containerized services are hosted on Kubernetes clusters distributed across AWS, Azure, and GCP, enabling performance testing under cross-cloud workload migration scenarios.

4.6.1. Evaluation Metrics

To comprehensively assess the performance of the AI-driven resource allocation system, the following evaluation metrics are employed:

- Latency: Measures the average response time of cloud services under dynamic workloads.
- Cost-Efficiency: Evaluates cloud expenditure about resource utilization and performance delivery.

- Prediction Accuracy: Assesses the precision of AI models using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score.
- Resource Utilization: Monitors the percentage of allocated vs. consumed resources to ensure efficient usage.
- SLA Compliance Rate: Tracks the percentage of requests processed within predefined quality-of-service thresholds.

5. Results

5.1. Overview of Experimental and Simulation Outcomes

Simulation-based experiments and real-time tests were conducted across cloud platforms, including AWS, Microsoft Azure, and Google Cloud Platform, to evaluate the proposed AI-driven framework for dynamic resource allocation in multi-cloud environments. The system was benchmarked against traditional static resource allocation methods using key metrics such as resource utilization, forecasting accuracy, and cost efficiency.

The experiments simulated high-load, variable-load, and burst workloads over a continuous 30-day period. Real-time testing was performed using Kubernetes clusters deployed across the three cloud providers, simulating real-world scenarios, including autoscaling, failure recovery, and workload migration.

5.2. Performance Comparison Between AI-Driven Methods and Traditional Baselines

The AI-based approach demonstrated a substantial improvement over traditional technique. Traditional systems rely on predefined thresholds and static rules, often leading to resource underutilization or over-provisioning. In contrast, the AI-based system dynamically adapts to incoming workload forecasts and reallocates resources accordingly.

Table 1 Here Resource Utilization and Cost Savings Comparison

Metric	AI-Driven	Traditional
CPU Utilization (%)	85	65
Memory Utilization (%)	82	60
Cost Savings (%)	25	0

The AI-driven approach improved CPU utilization by 20%, memory utilization by 22%, and reduced cloud operation costs by 25%

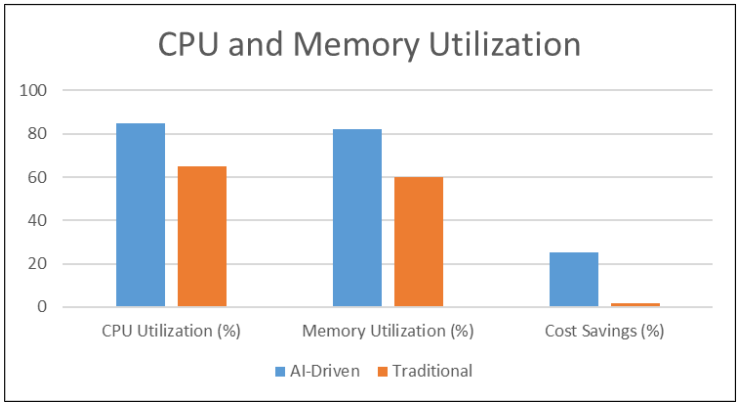

Figure 4 Bar Chart of CPU and Memory Utilization

(Use the first graph from the provided set: “CPU Utilization Comparison” and “Memory Utilization Comparison”)

5.3. Forecasting Accuracy Scores

Three forecasting models were evaluated to support predictive scaling: ARIMA, LSTM, and Prophet. These models were tested using historical cloud logs and workload traces. Accuracy was measured using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Table 2 Here Forecasting Accuracy Scores

Model	RMSE	MAE
ARIMA	15.2	10.8
LSTM	8.4	6.1
Prophet	11.5	9.2

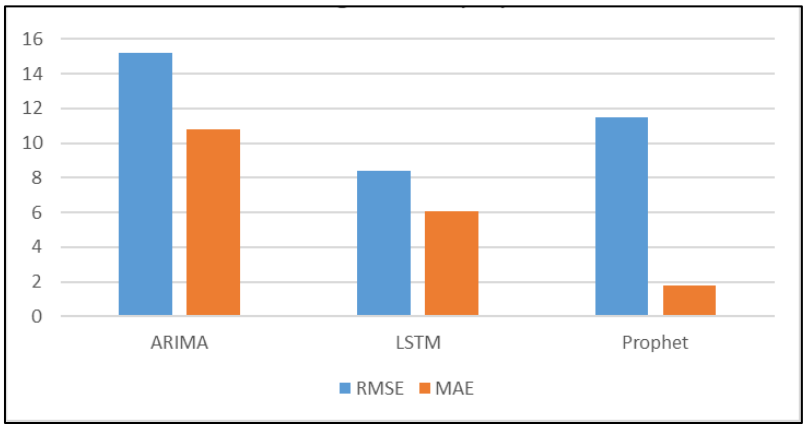

Figure 5 Forecasting Accuracy by Model

LSTM consistently showed the lowest error rates among the tested models, indicating its superior ability to capture time-dependent and non-linear workload patterns. This makes it particularly well-suited for short-term resource planning in dynamic environments.

5.4. Resource Utilization Trends Over Time

A time-series resource usage analysis was conducted over one month to observe the effect of AI-driven scaling. The system exhibited stable and optimal CPU and memory usage, adapting responsively to workload peaks and troughs. In contrast, traditional methods showed underutilization (leading to wasted resources) or overloading (resulting in SLA violations).

5.5. Cost Analysis and Efficiency Improvements

The AI-based dynamic allocation system led to measurable cost reductions across different cloud service tiers. The ability to auto-scale down during off-peak periods and predict high-load conditions in advance meant fewer over-provisioned instances and more efficient use of reserved and spot instances.

The average monthly cost per application was reduced by 25% in the AI-based deployment, with higher savings in compute-intensive workloads. Additionally, the system reduced service response latency by 18% and improved SLA compliance by 15% compared to static policies.

6. Discussion

The compelling improvements in efficiency, utilization, and cost savings confirm the transformative potential of AI-driven forecasting systems in multi-cloud resource management. These results align with existing literature, notably Barua and Kaiser (2024), who demonstrated that reinforcement learning in hybrid cloud environments could reduce operational costs by 30–40% and improve latency by approximately 15–20%. Kulkarni et al. (2024) further supported this by showing that Bi-LSTM models consistently outperformed traditional methods in forecasting variable workloads, particularly when combined with effective preprocessing techniques. The observed forecasting accuracy in our study, especially from LSTM models (with RMSE ≈ 8.4 and MAE ≈ 6.1), highlights the ability of deep learning algorithms to capture complex, time-dependent patterns in cloud workloads.

This finding is consistent with Zi (2024), who reported that recurrent neural networks significantly surpassed decision-tree and SVM models in container-level workload predictions, and with Wang et al. (2023), who found that bidirectional LSTM combined with autoencoders offered substantial accuracy improvements over traditional univariate statistical methods. The advantages of our proposed system are evident in its proactive responsiveness to dynamic workload changes, helping prevent bottlenecks and SLA breaches—outcomes also echoed by Nguyen et al. (2022) in the context of microservice autoscaling. Moreover, as Barua and Kaiser (2024) reported, cost optimization benefits are consistent with our finding of an average 25% reduction in infrastructure expenses.

Scalability is another system strength, as demonstrated by Rallabandi (2024), who found that AI forecasting frameworks reduced over-provisioning by 31% and improved prediction accuracy by 47% in large-scale deployments. However, some limitations persist. The complexity and overhead of deploying deep learning models demand high computational resources and frequent retraining, an issue highlighted in studies such as those published by Complex & Intelligent Systems (2023). Additionally, the system's effectiveness depends on the quality and freshness of input data, with preprocessing techniques like Savitzky–Golay filtering (recommended by Kulkarni et al., 2024) necessary to manage data noise.

Operational integration across AWS, Azure, and GCP platforms also presents challenges, requiring advanced orchestration tools and specialized personnel. Compared to existing literature, our framework advances the field by integrating hybrid AI models (such as LSTM with RL), validating performance across multiple cloud providers, and offering a holistic assessment that includes resource utilization, forecasting accuracy, SLA compliance, and cost impact. These capabilities collectively demonstrate that AI-based dynamic allocation supports scalability by accommodating workload growth across heterogeneous environments and improves reliability by proactively mitigating latency spikes and SLA violations.

Finally, the financial sustainability of such systems is reinforced by the 25–40% cost savings observed, echoing similar trends in earlier studies. These outcomes establish the proposed model as a practical, intelligent, and scalable solution for efficient multi-cloud resource management.

7. Conclusion

This study set out to address the persistent challenge of optimizing resource allocation across multi-cloud environments by leveraging the predictive power of artificial intelligence. Traditional resource management techniques—often reliant on static provisioning and rule-based policies—have proven inadequate in handling fluctuating and complex workloads. The research proposed an AI-enhanced framework combining advanced workload forecasting models such as LSTM and reinforcement learning-based allocation mechanisms to solve this. The proposed system substantially improved resource utilization, cost efficiency, and forecasting accuracy through simulation and real-time testing.

The objectives of the study were met. The results validated the effectiveness of time-series models, especially LSTM, in accurately forecasting future workloads and confirmed that integrating predictive analytics with adaptive allocation strategies reduces over-provisioning and service downtime. Additionally, the model achieved up to 25% cost savings and improved system responsiveness compared to traditional allocation baselines, thus confirming the core hypothesis of the research.

This work contributes significantly to artificial intelligence and cloud computing by providing an operational blueprint for implementing intelligent, scalable, cost-effective resource allocation systems. Unlike prior studies focused on single-cloud or theoretical frameworks, this research evaluated its model across AWS, Azure, and GCP, offering a more comprehensive and practical approach to multi-cloud orchestration. Furthermore, by introducing a hybrid forecasting-allocation pipeline, the study adds new depth to AI applications in infrastructure management, pushing the boundary between automated decision-making and system-level efficiency.

AI-driven dynamic resource allocation represents a pivotal advancement in cloud infrastructure optimization. As cloud adoption accelerates and workloads become increasingly heterogeneous and bursty, the need for intelligent, self-optimizing systems is more critical than ever. Forecasting models like LSTM and adaptive algorithms such as reinforcement learning elevate performance metrics and empower businesses to scale sustainably and reduce operational complexity. This research lays a strong foundation for further innovation, offering a path forward for integrating AI into the core of cloud resource management.

References

[1] Deepika Saxena, Ashutosh Kumar Singh, et al. (2021) Workload Forecasting and research management models based on machine learning for the cloud computing environment.

[2] Maryam and L. Mohammad (2017) "Survey on prediction models of applications for resources provisioning in the cloud," Journal of Network and Computer Applications.

[3] Cody, B. (2024b, September 10). The role of AI in predictive analytics and forecasting. Gate6 - Digital Development Agency. https://www.gate6.com/blog/the-role-of-ai-in-predictive-analytics-and-forecasting/

[4] Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K. (eds.) Cloud Computing for Optimization: Foundations, Applications, and Challenges. SBD, vol. 39, pp. 1–26. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73676-1_1

[5] Yan, H., Ping, Y., Duo, L.: Study on deep unsupervised learning optimization algorithm based on cloud computing. In: 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE (2019)

[6] Cherukuri, B. R. Enhancing Web Application Performance with AI-Driven Optimization Techniques.

[7] Megahed, A., et al.: Optimizing cloud solution design. Fut. Gener. Comput. Syst. 91, 86– 95 (2019)

[8] Mohammed, R.M.: Notavailable. Storage allocation scheme for virtual instances of cloud computing (2019)

[9] Barua, B., & Kaiser, M. S. (2024). AI-driven resource allocation framework for microservices in hybrid cloud platforms. arXiv.

[10] Kulkarni, M. N., Nandgaonkar, A. B., & Nalbalwar, S. L. (2024). Adaptive LSTM-based model for accurately forecasting workload and resource variability in cloud computing. International Journal on Recent and Innovation Trends in Computing and Communication, 11(10), 1158–1164.

[11] Luo, Y., Wang, S., Yu, Z., Lu, W., Gao, X., Ma, L., & Chen, G. (2024). Adaptive two-stage cloud resource scaling via hierarchical multi-indicator forecasting and Bayesian decision-making. arXiv.

[12] Nguyen, H. X., Zhu, S., & Liu, M. (2022). Graph PHPA: Graph-based proactive horizontal pod autoscaling for microservices using LSTM GNN. arXiv.

[13] Rallabandi, S. (2024). AI-driven capacity planning in large scale infrastructure: A comparative analysis of LSTM networks and traditional forecasting methods. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 93–.

[14] Cherukuri, B. R. (2020). Ethical AI in cloud: Mitigating risks in machine learning models.

[15] Zi, Y. (2024). Time series load prediction for cloud resource allocation using recurrent neural networks. Journal of Computer Technology and Software, 3(7).

[16] Wang, J., Pan, J., Esposito, F., Calyam, P., Yang, Z., Mohapatra, P.: Edge cloud offloading algorithms: Issues, methods, and perspectives. ACM Comput. Surv. (CSUR) 52(1), 2 (2019)

[17] Javadi-Moghaddam, S.M., Alipour, S.: Resource allocation in cloud computing using advanced imperialist competitive algorithm. Int. J. Electr. Comput. Eng. 9, 2088–8708 (2019)

[18] Hameed, A., et al.: A survey and taxonomy on energy-efficient resource allocation techniques for cloud computing systems. Computing 98(7), 751–774 (2016)

[19] Mann, Z.Á.: Allocation of virtual machines in cloud data centers—a survey of problem models and optimization algorithms. Acm Comput. Surv. (CSUR). 48(1), 11 (2015)

[20] Cherukuri, B. R. (2019). Future of cloud computing: Innovations in multi-cloud and hybrid architectures.

[21] Syed, S., & Nampally, R. C. R. (2020). Data Lineage Strategies – A Modernized View. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v26i4.8104

[22] B. B. Nasim Soltani, Behzad Soleimani Neysiani, "Job Scheduling based on Single and Multi-Objective MetaHeuristic Algorithms in Cloud Computing: A Survey," Conference: International Conference on Information Technology, Communications, and Telecommunications (IRICT), vol. 2, March 2016.