iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83704

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Cost-Optimized Cloud Resource Allocation Using Usage Patterns

Mr. Shravan Deepak Talegaonkar, Dr. Suhas Rautmare
Dept. of Information Technology University of Mumbai Mumbai, India

Abstract: ***Cloud computing provides scalable computing resources which are crucial for modern organizations. It should be noted that efficient resource management is challenging due to differences in workload characteristics. There exist certain resource management techniques which utilize scaling and threshold methods. At the same time, they are characterized by inefficiency in terms of resource usage and operation cost. Machine learning, deep learning and reinforcement learning techniques provide other alternatives, but they demand great computation capacity and expenses. Therefore, this research provides for allocation of resources based on detection of their usage patterns through application of machine learning algorithms. Usage patterns are detected based on historical CPU and memory usages data along with trend analysis. The proposed method makes intelligent allocation of cloud computing resources possible thanks to evaluation of thresholds in combination with trend analysis. An extensive literature review concerning the existing solutions for cloud resource optimization was conducted in order to detect current problems and identify areas of further study. The developed technique may be integrated into the monitoring services offered by various clouds including AWS CloudWatch in order to ensure intelligent management of cloud computing resources.***

I. INTRODUCTION

One of the significant technologies that have been implemented in modern information technology systems is cloud computing. This involves using the internet to have access to different computing resources on demand. Rather than building expensive IT infrastructure, it becomes easy for firms to use cloud services where they can access computing resources such as virtual computers, storage, software, and networking at any point in time. Due to its efficiency, affordability, and scalability, cloud computing has been embraced widely in various industries including business, health care, educational institutions, financial sector, and governments.

With more firms relying on cloud computing, efficient management of cloud computing resources has emerged as one of the key challenges faced by many cloud service providers. Allocating computing resources to different user applications becomes an important issue as cloud computing resource providers strive to maintain high levels of performance while keeping their expenses low. However, the demands of workloads in cloud computing environment tend to be very dynamic in nature, thus changing over time. Unexpectedly, there could be a rise in demand for cloud computing services due to increased activities by users.

Allocation of resources involves the distribution of existing cloud computing resources among applications based on their needs. The most important thing is to balance the trade-off between performance and cost. Allocation of many resources would make provision overprovisioning and hence cause waste of resources and unnecessary cloud costs. On the other hand, allocation of few resources would create underprovisioning and could affect the performance of cloud computing services negatively due to slower response time, service disruptions, and low performance. Therefore, it is critical to manage cloud computing resources efficiently to maximize performance while minimizing costs.

Most of the cloud resource management systems are reactive and use static threshold-based rules. When resource usage exceeds some pre-specified levels, resources are added, whereas reduction occurs once resource usage falls below certain levels. Though these systems are easy to implement and popular, they are not efficient to handle fast-changing workload since the decision on scaling up or down can be taken only after the actual change in resource utilization. This makes reactive systems inefficient and costly to operate.

In order to address such shortcomings, a variety of intelligent approaches for managing resource requirements have been suggested. These include the application of machine learning, prediction, deep learning, and reinforcement learning to automate and optimize resource allocation decisions. Based on the analysis of thirty-nine research papers, one can identify a number of significant achievements in such areas as predictive resource allocation, workload forecasting, minimizing costs for cloud resources, auto-scaling algorithms, reinforcement learning for task scheduling, and multi-cloud resource management, among others.

©IJRASET: All Rights are Reserved

Although these algorithms contribute to decision-making and resource allocation, they usually need to work with large sets of training data, advanced models, significant computing power, and regular maintenance.

One of the main observations that could be made about the behavior of applications hosted on cloud resources is the existence of periodic consumption patterns in business applications. For instance, there would always be more active users of applications during business hours than outside of those hours. This holds true for some cloud computing services as well. In such cases, it is possible to predict the workload using previously collected data and the knowledge about usage patterns without involving sophisticated machine learning models.

The presented research aims at designing an efficient and cost-effective mechanism for cloud resource allocation based on analyzing workload patterns and making data-driven decisions. Specifically, the proposed solution leverages historical data related to CPU and RAM workload that is provided by various sources like CSV files and monitoring systems such as AWS CloudWatch. The collected data is then timestamped and grouped according to time to uncover workload patterns. On the basis of workload trends, the system is able to establish whether there is an increase, decrease, or stability in utilization and make the proper scaling decisions based on priority rules.

In contrast to threshold-based mechanisms for resource allocation, the proposed approach uses workload trends when making decisions. On the other hand, similarly to machine learning and reinforcement learning models, the system does not need any model training as well as large amounts of data or significant computing resources.

The suggested framework further ensures cost awareness as resource allocation decisions are assessed from both technical and cost perspectives. Based on the comparison of optimized resource allocation approaches with traditional allocation methods, the system is capable of making cost estimations while ensuring acceptable performance levels. It allows businesses to assess their cloud infrastructure and spending more precisely.

The key benefits of the research include the implementation of historical usage pattern analysis for resource allocation purposes, workload trends identification, scaling recommendations based on predictions, lightweight rule-based decision engine, integration with AWS CloudWatch metrics, and cloud cost optimization without resorting to advanced machine learning algorithms. The suggested approach focuses on providing a realistic, scalable, and deployable model for optimizing cloud resources.

It was shown through the study conducted that intelligent optimization of cloud resource utilization can be achieved by analyzing the historical behavior of workloads based on pattern identification without the need for sophisticated artificial intelligence systems for resource management.

II. BACKGROUND

A. Reactive vs Proactive Resource Allocation

Resource allocation systems of traditional clouds predominantly run on reactive processes. Such systems keep monitoring the level of utilization of resources such as CPUs, memories, and networks. Once this resource utilization reaches the pre-specified levels, more resources will be made available to keep applications running at their best. For instance, if CPU utilization increases beyond 80 percent, then there is automatic scaling for additional CPU resources.

Although reactive systems are easy and popular, the drawback associated with them is that they take time to react to changes because scaling takes place after resource utilization reaches a particular level. During such delays, users can face problems like poor performance or even interruption of services. This problem becomes acute when there is a sharp rise in traffic volume.

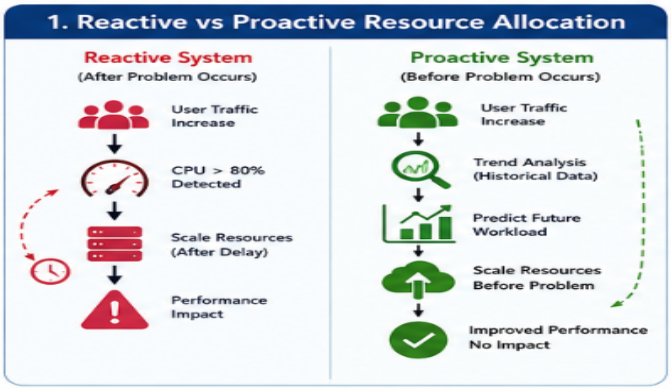


Figure 1 illustrates the difference between reactive and proactive resource allocation mechanisms, highlighting how trend-based analysis enables earlier and more effective scaling decisions.

The proactive approach shown in Figure 1 addresses this limitation by analyzing historical workload behavior and identifying workload trends before resource exhaustion occurs. Instead of waiting for threshold violations, the system predicts future workload growth using trend analysis and initiates scaling actions in advance. As a result, proactive resource allocation improves application performance, reduces service disruption, and enhances user experience.

B. Historical Data-Based Analysis

Examples of cloud monitoring solutions are the AWS CloudWatch platform which gathers operational data on resource usage. Some of the metrics collected include CPU usage, memory usage, network usage, and disk operations. The use of monitoring data helps in getting an understanding of the characteristics of the workload as well as how resource usage patterns occur over time.

Most resource allocation strategies only focus on current usage data but neglect historic data. This has the effect of making any emerging workload patterns difficult to recognize, hence poor decision-making.

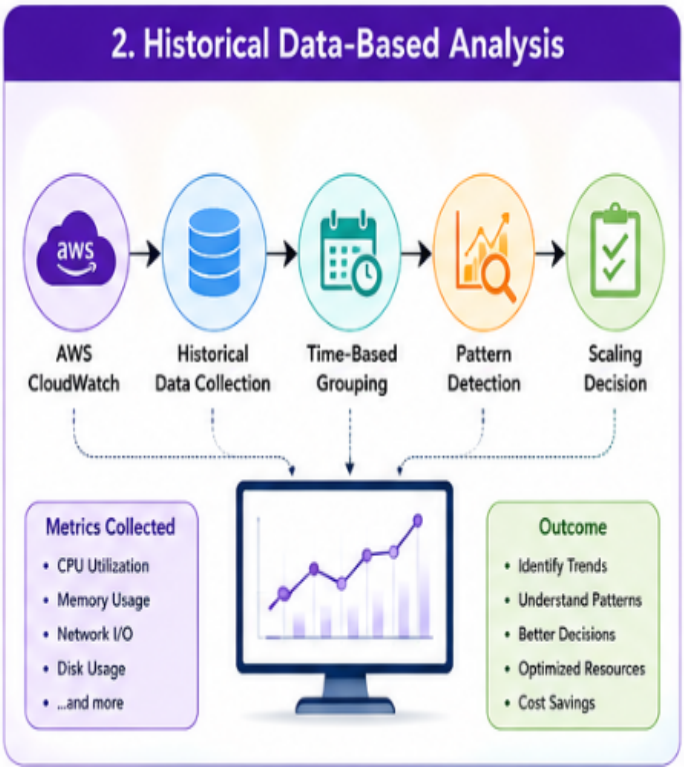


Figure 2 presents the workflow of historical data analysis, beginning with AWS CloudWatch data collection and ending with pattern detection and scaling decision generation.

The proposed framework utilizes historical resource utilization data collected from cloud monitoring services. The collected data is organized according to timestamps and divided into meaningful time intervals. Pattern detection techniques are then applied to identify workload behavior and generate scaling recommendations. By leveraging historical information, the system can better understand workload characteristics and support proactive resource management.

C. Traditional vs Enhanced Rule Engine

Rule-based decision-making systems are some of the most prevalent types of methods employed in the management of resources within clouds. Classic rule engines operate using thresholds that trigger particular scaling operations. Resources can be scaled up when CPU usage goes above a certain threshold and scaled down once usage goes below a particular threshold.

Despite their ease of deployment and low demand on computing capacity, these systems have no notion of context. Resource-scaling operations are only driven by current levels of resource utilization with no regard for trends in utilization. As a result, the same utilization rates may require different amounts of resources depending on the trend of usage.

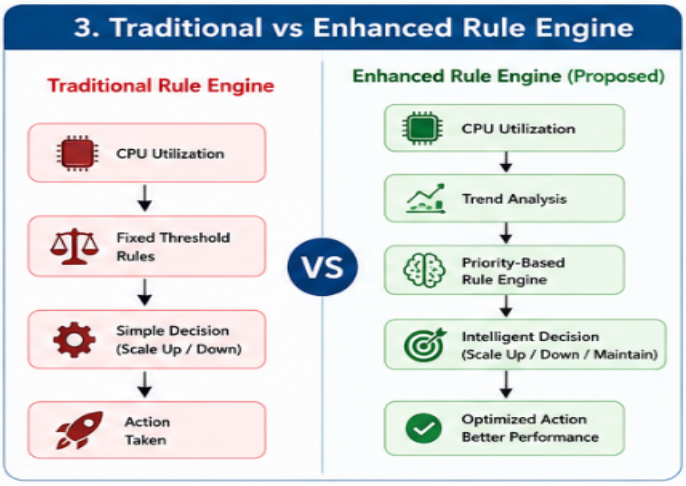


Figure 3 compares the traditional threshold-based rule engine with the proposed enhanced rule engine that incorporates workload trend analysis and priority-based decision-making.

The proposed framework introduces an enhanced rule engine that combines utilization metrics with trend analysis. Before making a scaling decision, the system evaluates workload patterns and determines whether resource demand is rising, stable, or declining. A priority-based rule engine then generates the most appropriate scaling recommendation. This approach improves decision accuracy while maintaining the simplicity of rule-based systems.

D. Machine Learning vs Proposed Approach

Recently, the application of machine learning, deep learning, and artificial intelligence has been proposed for resource management in cloud computing environments. These methods employ historical workload information and construct predictive models that predict future resource demands. They tend to be highly accurate predictors and help automate decision-making processes.

However, while possessing considerable benefits, machine learning-based solutions need sizable training sets, substantial computational power, and continual model tuning. The use of such methods can lead to higher costs and greater complexities in terms of managing IT infrastructure. Moreover, small firms may find it challenging to adopt such tools.

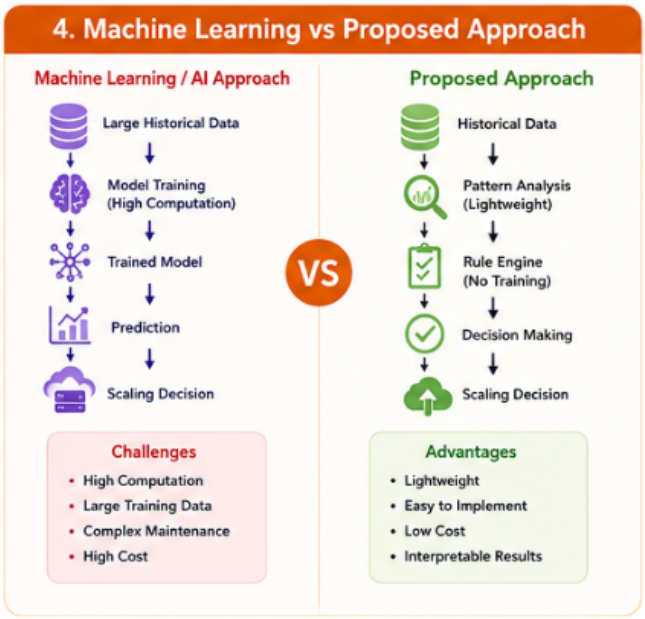


Figure 4 highlights the differences between machine learning-based resource allocation and the proposed pattern-based approach, emphasizing simplicity, lower cost, and ease of implementation.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

The proposed framework follows a lightweight alternative approach. Instead of training predictive models, it directly analyzes historical workload patterns and trends. Resource allocation decisions are generated using a rule-based decision engine that operates without model training. This significantly reduces computational overhead while maintaining effective resource optimization capabilities.

E. Pattern-Based Resource Allocation

In pattern based resource allocation, it is about analyzing workload behavior throughout its lifecycle instead of using only the value of resource usage at one point in time. The historical records of resource utilization can have a number of recurring patterns that describe how workloads behave with respect to their consumption of resources.

The pattern-based approach utilizes historical data about CPU and memory utilization to analyze the following types of workloads: increasing workload pattern, steady workload pattern, and decreasing workload pattern. An increasing workload pattern suggests an increase in workload behavior while a decreasing workload pattern suggests the reverse behavior.

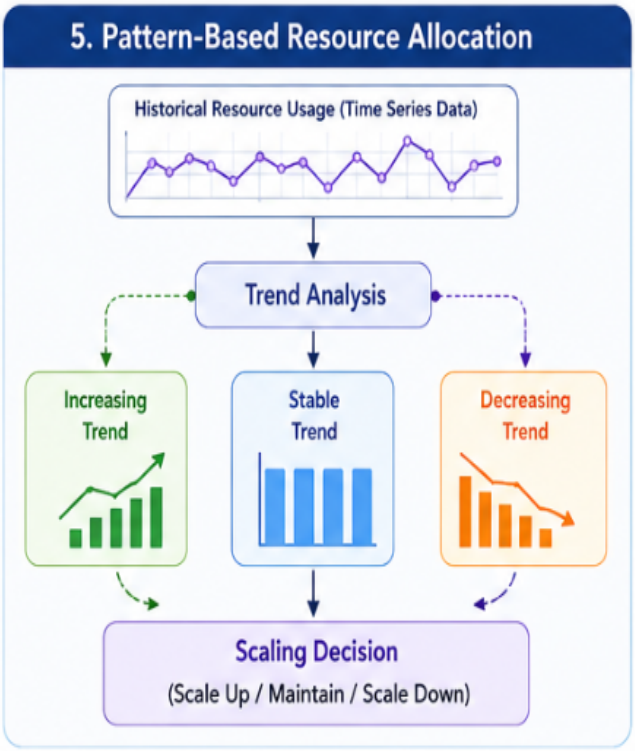


Figure 5 demonstrates the process of pattern classification and illustrates how trend analysis supports intelligent scaling decisions based on workload behavior.

By classifying workload behavior into these categories, the system gains a deeper understanding of future resource requirements. This enables more informed scaling decisions compared to traditional threshold-based approaches. Pattern-based allocation also remains computationally efficient because it does not require complex machine learning models.

F. Cost-Aware Cloud Resource Optimization

One of the most important aspects that should be considered while implementing cloud computing is cost efficiency since the payment is done depending on the use of resources. Failure to allocate resources efficiently might cause unnecessary high costs or underperforming applications. It becomes essential to balance resource usage, application performance, and economic factors.

The suggested approach involves cost-awareness in the process of allocating resources. Initially, resource usage is evaluated for recognizing workload characteristics and identifying measures required to allocate necessary resources. Then, the proposed method is evaluated economically to recognize the benefits achieved by implementing the proposed approach.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com



Figure 6 illustrates the complete cost optimization cycle, showing how monitoring, pattern analysis, scaling decisions, cost evaluation, and optimized allocation work together to achieve efficient cloud resource management.

By reducing unnecessary resource provisioning while maintaining acceptable performance levels, the framework helps organizations optimize cloud spending. This approach supports both technical and business objectives by improving resource utilization and minimizing infrastructure costs.

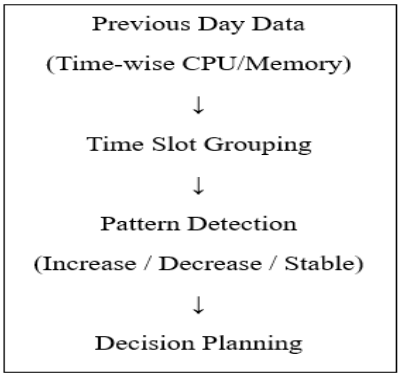
III. CLASSIFICATION OF MECHANISM

A. Pattern-Based Mechanism

The system is classified as a pattern-based mechanism because it focuses on how workload behaves over time using past data.
It identifies patterns such as:

- Increasing trend
- Decreasing trend
- Stable trend

These patterns are extracted from previous-day time slots, and decisions are made based on these observed behaviors.
Example:
If CPU usage from 9–10 AM to 10–11 AM shows a continuous increase in the previous day, the system assumes a similar pattern and plans scaling accordingly.



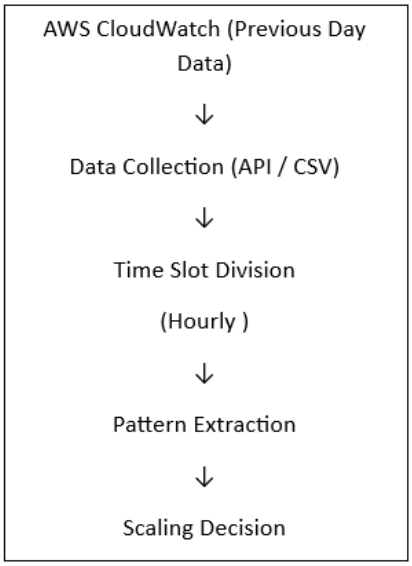
B. Historical Data-Driven Approach

The system is fully based on historical data, specifically previous-day usage collected from sources like AWS CloudWatch.
This data is:

©IJRASET: All Rights are Reserved

- Organized using timestamps
- Divided into time slots
- Analyzed to identify repeating patterns

Decisions are derived from past behavior.

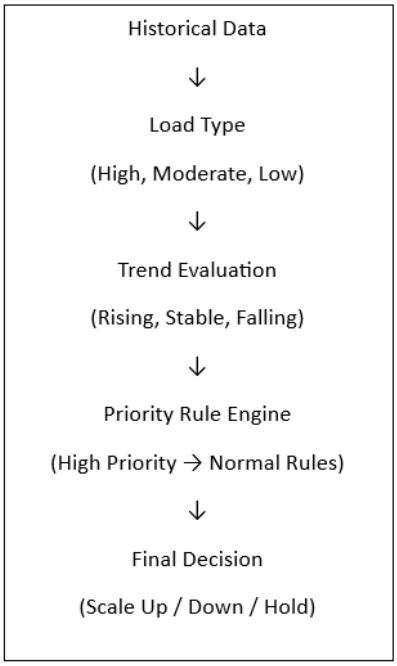


C. Rule-Based Decision System (Refined)

The system uses predefined rules, but these rules are not simple or isolated. They work in a priority-based structure and use both:

- Patterns / Load Type (High, Moderate, Low)
- Trend (Rising, Stable, Falling)over time

So, decisions are not made by a single condition, but by a combination of rules + pattern+trend evaluation.



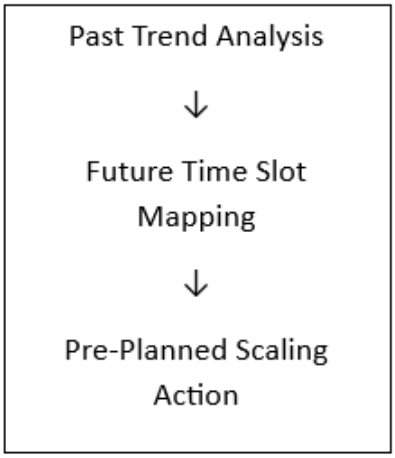
D. Proactive (Trend-Based Planning Mechanism)

It is proactive suggestion system.

It prepares scaling decisions in advance based on previously observed trends.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Example:
If every day at 10 AM usage increases, the system plans scaling before that time arrives.
This avoids last-minute decisions and ensures smoother performance.



E. Resource Allocation Method

The system updates decisions periodically based on newly available historical data.

- It is not static (fixed forever)
- It is not fully real-time

Instead, it:

- Uses previous-day data
- Updates decisions for upcoming time slots
- Adapts over time



IV. LITERATURE REVIEW

A. Traditional Resource Allocation and Cost Optimization Approaches

Resource Allocation in Cloud Computing Resource allocation in cloud computing presents a major challenge to the cloud service providers in maintaining the performance, availability, and the cost of providing their services. Various conventional resource allocation approaches include static provisioning, threshold based scalability, heuristic algorithms, and optimization techniques which try to optimize resource allocation and minimize the cost incurred by the providers.

Nagalla et al. [1] examined adaptive resource allocation and cloud cost optimization techniques showing that efficient resource allocation minimizes costs of service provision while ensuring quality delivery.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Also, Deochake [13] reviewed several cloud cost optimization methods and emphasized the importance of allocating cloud resources in an adaptive manner, resource rightsizing and tracking of utilization to cut cloud costs. Nwanganga et al. [20] proposed minimum cost flow approach to optimize workloads and showed that using optimization techniques to allocate computing resources can minimize infrastructure costs. Swain et al. [21] examined efficient resource management approaches in cloud computing and emphasized the importance of virtualization, workload balancing and resource scheduling. Mukherjee et al. [22] gave an overview of various cloud resource management methodologies and noted the significance of resource allocation in cloud computing.

Additionally, Pingulkar et al. [25] reviewed traditional resource allocation strategies such as static, dynamic, market-based, and heuristic-based resource allocations. They highlighted the fact that even though traditional strategies are straightforward and computationally efficient, they have been shown to be ineffective in managing dynamic workloads.

These studies collectively indicate that while traditional resource allocation techniques provide an initial approach to cloud computing, they lack in terms of flexibility and adaptability to handle dynamic workloads.

B. Machine Learning-Based Resource Allocation

Machine learning algorithms have been found to be useful as a potential solution to the problem of cloud resource allocation, where such systems learn from their past data to make smart allocation decisions.

Wang & Yang [2] presented a resource allocation approach using machine learning algorithms that would optimize the cloud resources allocation by making smarter decisions. Cesarini [4] discussed various machine learning algorithms used for cloud resource allocation and found out that machine learning based predictions result in more efficient allocation of cloud resources than any static allocation approach.

Killedar et al. [5] designed a cost optimization framework for the cloud using machine learning techniques to assign appropriate cloud resources based on workload behavior. Zhang et al. [9] analyzed the role of machine learning optimization in cloud resource scheduling. Yadav and Yadav [17] explained that predictive resource management using machine learning reduces resource wastage without compromising performance.

Anbarkhan [23] evaluated multiple machine learning methods used in cloud resource optimization and pointed to the significance of employing intelligent computing methods to handle cloud scalability issues in the present day. Overall, from the research performed, it is clear that machine learning allows better resource allocation, taking advantage of past workload data and prediction algorithms.

Unfortunately, most machine learning applications involve large datasets and a significant amount of computation, as well as ongoing maintenance.

C. Predictive Analytics and Workload Forecasting

Workload prediction is becoming an increasingly crucial area of study in cloud resource management as it facilitates better cost management through resource planning for upcoming workloads.

For example, Smith [3] designed workload scheduling based on predictive analytics with an objective to minimize cloud resource cost by using workload forecasting. Similarly, Kamble et al. [7] used workload prediction in cloud environment with the help of machine learning for efficient resource management. Chanthati [8] studied cloud resource planning using predictive analytics and pointed out the significance of forecasting for unnecessary resource management.

Furthermore, Saxena and Singh [11] used workload forecasting models along with machine learning algorithms for better management of cloud resources. In another study, Zheng et al. [12] used AI-based predictive analytics for resource allocation with enhanced accuracy than traditional methods.

Moreover, Smendowski and Nawrocki [16] worked on multi time series-based workload utilization forecasting. Rossi et al. [19] also contributed toward improving workload forecasting by incorporating uncertainty awareness and transfer learning techniques.

A framework to predict resource utilization using machine learning and evolutionary algorithms was suggested by Malik et al. [29]. On the other hand, Temporal Fusion Transformers (TFT) offered by Lim et al. [30] are an interpretable method for predicting data on multiple horizons, thus gaining importance in cloud workload prediction.

Predictive auto-scaling and workload prediction were the topics of research conducted by Roy et al. [35] and Shariffdeen et al. [36], respectively, proving the significance of precise predictions in terms of resource allocation optimization and cost minimization.

Thus, predictive analytics offers proactivity in cloud management through anticipation of future requirements.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. AI-Driven Cloud Resource Optimization

The advent of Artificial Intelligence has not only broadened the scope of cloud resource management from simple forecasting and machine learning but has made intelligent decision-making possible in dynamically evolving cloud platforms.

Chandrakanth [15] proposed an AI-based dynamic resource allocation framework that incorporated both predictive modeling and real-time optimization processes. The study proved that AI-based techniques could effectively modify resource allocation depending on dynamically changing workloads.

Similarly, Ramamoorthi [24] introduced an AI-based framework for the optimization of resources in cloud computing through automated resource allocation. He discussed the importance of intelligent automation for efficient resource management in the cloud.

Both these studies show how AI-based optimization frameworks have improved the adaptability and automation aspect of cloud computing. However, complexities related to implementation and increased computational cost are still a serious concern.

E. Reinforcement Learning-Based Resource Allocation

Reinforcement Learning (RL) represents an innovative resource management tool with the capability to learn optimal resource allocation policy through interactions with cloud systems over time.

In their paper, Kayalvili et al. [14] implemented reinforcement learning for resource allocation in clouds and obtained increased resource utilization efficiency and allocation efficiency. On the other hand, reinforcement learning was employed in the dynamic resource allocation problem in a multi-cloud environment by Varghese [28], resulting in better decision-making.

Deep reinforcement learning was applied to the cloud resource management domain by Mao et al. [31], illustrating the benefits of such algorithms in resource allocation. In another study, deep reinforcement learning was implemented in the context of cloud resource scheduling by Ye et al. [32], obtaining better performance than classical algorithms.

Liu et al. [33] developed a hierarchical model for resource management and energy optimization in clouds based on deep reinforcement learning. This algorithm showed great success in terms of both energy consumption and performance.

It is clear from these studies that reinforcement learning provides many opportunities for adaptive cloud management, although considerable training periods are required.

F. Reinforcement Learning Surveys and Research Trends

With the rising adoption of reinforcement learning in cloud computing, various researchers have conducted extensive literature reviews to evaluate its efficiency and possible future applications.

Garí et al. [34] reviewed reinforcement learning approaches for auto-scaling in cloud computing and reported substantial advancements in resource elasticity and application performances. Zhou et al. [37] presented a detailed literature review on deep reinforcement learning for cloud resource management and scheduling and mentioned future directions in intelligent cloud computing.

Gu et al. [38] presented a detailed literature review at the algorithm level for deep reinforcement learning in cloud computing for resource scheduling and resource management. The study focused on the increasing use of RL approaches in cloud computing along with associated issues of scalability, convergence, and computational complexity.

The above literature reviews indicate that reinforcement learning will play an important role in future autonomous cloud management systems.

G. Multi-Cloud, Hybrid Cloud and Edge Resource Management

Cloud environments have evolved to include many more cloud service providers and even edge computing environments, posing fresh resource allocation problems.

A framework using artificial intelligence was suggested by Barua and Kaiser [18] for microservice-oriented resource allocation in hybrid cloud systems. Their method offered efficient resource allocation without compromising on application performance.

Similarly, Sinha [26] explored multi-cloud workload placement and found that effective workload management was key to improved performance and cost-efficiency. Chen [27] looked at resource management in the cloud-edge context and noted the importance of intelligent resource management in such a scenario.

Finally, Varghese [28] showed the efficacy of reinforcement learning for resource management in the case of multi-clouds. Clearly, the evolving infrastructure requires solutions capable of dealing with these scenarios.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

H. Auto-Scaling and Dynamic Resource Scheduling

Proactive auto-scaling continues to be an essential approach to sustaining high-quality application performance without resource wastage.

A predictive approach to auto-scaling using workload predictions has been presented by Roy et al. [35]. Similarly, Shariffdeen et al. [36] have introduced adaptive approaches to workload predictions to enable proactive auto-scaling in Platform-as-a-Service environments.

Reinforcement learning-based proactive auto-scaling approaches have been comprehensively reviewed by Garí et al. [34], and modern approaches to cloud auto-scaling have been reviewed by Alharthi et al. [39].

From the literature review, it is evident that proactive and intelligent auto-scaling approaches outperform traditional threshold approaches in terms of low response times and effective resource usage.

I. Comparative Analysis and Research Observations

Several surveys and review articles have tried to compare the various resource allocation mechanisms and pinpoint any gaps in the literature.

For instance, Bodra and Khairnar [6] offered a comparative analysis of the use of machine learning algorithms in resource allocation techniques and noted that the major trade-off involved was in terms of accuracy versus computational efficiency. Similarly, Khan et al. [10] provided an extensive review of cloud resource management algorithms using machine learning techniques and pinpointed the future direction for intelligent cloud systems.

Additionally, Deochake [13], Pingulkar et al. [25], Garí et al. [34], Zhou et al. [37], Gu et al. [38], and Alharthi et al. [39] have all made it clear that while both machine learning and reinforcement learning algorithms offer more efficient allocation of resources, they tend to be computationally expensive, requiring huge amounts of data and complex implementation mechanisms.

The literature has clearly indicated that there is a huge gap between highly complicated AI-based techniques and very simplistic threshold-based mechanisms. The majority of the existing techniques are either too unintelligent or overly complicated. This gap calls for the development of intelligent resource allocation methods based on user behavior.

Paper	Author & Year	Approach Used	Key Contribution	Limitation
1	Nagalla et al. (2025)	Adaptive Resource Allocation	Improves cost efficiency using dynamic allocation	Reactive, lacks prediction
2	Wang & Yang (2025)	Machine Learning	Intelligent allocation using ML algorithms	High complexity and computation
3	Smith (2025)	Predictive Analytics + ML	Cost-aware workload scheduling	Requires continuous model training
4	Cesarini (2024)	Machine Learning	Improves resource utilization using ML	No practical implementation
5	Killedar et al. (2025)	ML-based Optimization	Reduces cost using usage data	Requires training and resources
6	Bodra & Khairnar (2025)	Comparative ML Review	Compares ML algorithms for cloud	High complexity in hybrid models
7	Kamble et al. (2023)	Predictive ML	Forecasts workload for scaling	Depends on accurate historical data
8	Chanthati (2025)	Predictive Analytics	Cost forecasting and planning	No real-time allocation
9	Zhang et al. (2023)	ML Optimization	Efficient resource scheduling	High computational cost
10	Khan et al. (2021)	ML Review	Overview of ML-based resource systems	Generalized, lacks practical solution
11	Saxena & Singh (2021)	ML Forecasting	Predicts workload demand	Does not optimize cost directly
12	Zheng et al. (2024)	AI Predictive System	Improves allocation efficiency	Data-dependent and complex
13	Deochake (2023)	Cost Optimization Review	Reviews cost-saving strategies	No intelligent system
14	Kayalvili et al. (2025)	Reinforcement Learning	Dynamic resource allocation using prediction-enabled RL	Very high complexity

15	Chandrakanth (2024)	AI-Based Allocation	Real-time adaptive resource optimization	High implementation cost
16	Smendowski & Nawrocki (2024)	Time-Series Forecasting	Multi-series workload prediction	Computational overhead
17	Yadav & Yadav (2025)	ML-Based Predictive Management	Improves resource utilization and forecasting	Requires extensive training data
18	Barua & Kaiser (2024)	AI-Driven Hybrid Cloud Framework	Optimizes microservice resource allocation	Complex hybrid architecture
19	Rossi et al. (2023)	Workload Forecasting + Transfer Learning	Uncertainty-aware workload prediction	Forecasting accuracy depends on data quality
20	Nwanganga et al. (2017)	Minimum-Cost Flow Model	Cost-efficient workload optimization	Limited adaptability to dynamic workloads
21	Swain et al. (2022)	Resource Management Framework	Improves cloud resource utilization	Lacks predictive capabilities
22	Mukherjee et al. (2024)	Resource Management Survey	Comprehensive overview of cloud resource management	Theoretical focus
23	Anbarkhan (2025)	ML-Based Resource Optimization	Efficient cloud resource allocation using ML	Computationally expensive
24	Ramamoorthi (2021)	AI-Driven Optimization Framework	Real-time resource allocation decisions	High deployment complexity
25	Pingulkar et al. (2023)	Resource Allocation Review	Comprehensive analysis of allocation techniques	No practical implementation
26	Sinha (2024)	Multi-Cloud Optimization	Optimizes workload placement across clouds	Limited real-world validation
27	Chen (2021)	Cloud-Edge Resource Management	Intelligent optimization in cloud-edge environments	High coordination complexity
28	Varghese (2024)	Reinforcement Learning	Dynamic allocation in multi-cloud environments	Long training time
29	Malik et al. (2022)	ML + Evolutionary Algorithms	Resource utilization prediction	Requires large historical datasets
30	Lim et al. (2021)	Temporal Fusion Transformer (TFT)	Interpretable multi-horizon forecasting	Computationally intensive
31	Mao et al. (2016)	Deep Reinforcement Learning	Automated cloud resource management	Training instability
32	Ye et al. (2018)	Deep RL Scheduling	Intelligent resource scheduling	High computational requirements
33	Liu et al. (2017)	Hierarchical Deep RL	Joint resource allocation and power management	Complex architecture
34	Garí et al. (2020)	RL-Based Auto-Scaling Survey	Comprehensive review of RL autoscaling	No implementation framework
35	Roy et al. (2011)	Predictive Auto-Scaling	Forecast-based scaling decisions	Prediction errors affect performance
36	Shariffdeen et al. (2016)	Adaptive Workload Prediction	Proactive auto-scaling for PaaS systems	Limited scalability evaluation
37	Zhou et al. (2024)	DRL Review	Review of DRL resource scheduling methods	Survey-only contribution
38	Gu et al. (2025)	DRL Algorithm Review	Analysis of DRL algorithms for cloud management	Lacks experimental validation

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

39	Alharthi et al. (2024)	Auto-Scaling Survey	Identifies challenges and future directions in auto-scaling	No proposed solution

V. RESEARCH GAP

Traditional vs Usage pattern(Data driven)

Aspect	Traditional Approach	My Approach
Strategy	Reactive	Pattern-based proactive
Data Usage	Current only	Historical + trend analysis
Prediction	None	Based on patterns
Complexity	Low	Moderate
Cost Efficiency	Low	Improved
Implementation	Simple rules	Analytical logic
Intelligence Level	Static	Data-driven

Static traditional approach includes:

- CPU > 80% → Scale up
- CPU < 20% → Scale down

Data Driven new approach includes:

"Data-driven intelligence refers to decision-making based on analysis of historical data and usage patterns, allowing the system to adapt to changing workload behavior."

Case 1:

- CPU = 70% (not very high)
- But trend = increasing

New system:

"Load is rising → prepare to scale up"

Case 2:

- CPU = 70%
- But Trend = decreasing

New system:

"Load is dropping → no need to scale"

A. Over-Reliance on Reactive Scaling

Most of the current cloud resource allocation schemes use a reactive threshold-based approach whereby scaling is done after resource usage breaches certain levels. This can be seen in research work by Nagalla et al. [1], as well as that done by S. Smith [3]. Reactive approaches tend to respond late since they react after workloads have begun affecting performance. This causes latency or degradation of services for a period of time before scaling takes effect.

In addition, the lack of anticipation of future workloads results in inefficient scaling.

B. Underutilization of Historical Pattern Behavior

Many scholars, like Wang and Yang [2] and Killedar et al. [5], make use of historical information for workload prediction. Their methodologies generally consider statistics averages or even short-term predictions, without studying the behavior pattern of workloads over time.

Their method does not account for whether the workload has been increasing, decreasing, or changing during certain periods of time. Thus, they do not gain valuable insight regarding workload changes.

Inadequate pattern analysis based on historical trends leads to inaccurate decisions.

C. *High Complexity of Machine Learning Models*

Several sophisticated methods that use techniques like machine learning and reinforcement learning have been proposed by researchers such as Cesarini [4], Bodra& Khairnar [6], and Khan et al. [10].
Even though they perform well, these methods suffer from certain limitations:

- Necessity of huge data sets to train the model
- High computational cost
- Extensive complexity to implement and tune the model
- Unavailability to small companies due to their complexity

Machine learning-based systems tend to be too complex to deploy effectively.

Aspect	Existing Research (ML)	My Approach
Approach	ML / Deep Learning	Pattern-based rules
Complexity	High	Low
Training Required	Yes	No
Computational Cost	High	Low
Implementation	Difficult	Easy
Practical Use	Limited	High

D. *Lack of Practical Implementation Feasibility*

Some works, such as Chanthati [8] and Saxena & Singh [11], emphasize simulation-oriented models or theories. While these models help in gaining academic insights, they do not account for practical implementation issues.
The issues:

- Challenge in deploying to actual clouds
- Insufficient emphasis on API-driven approa
- Failure to validate using actual data

E. *Absence of Cost-Focused Decision Integration*

For example, Killedar et al. [5] and Zhang et al. [9] consider performance measures, such as the response time and the CPU usage, but there is no explicit focus on cost optimization in their approaches to scaling.
Cost is important when we talk about scaling issues because any inefficiency in scaling may cost money.
Cost-related lack of awareness of scaling algorithms.

F. *Lack of Explainable Resource Allocation Decisions*

Several sophisticated resource allocation strategies depend on the use of machine learning, deep learning, and reinforcement learning methods for scaling decisions. Cesarini [4], Kayalvili et al. [14], Mao et al. [31], and Ye et al. [32] examine intelligent resource allocation using automatic learning models.
While these methods offer great predictive precision, the underlying decision process remains opaque. Cloud operators might receive scaling suggestions but have no clue about the criteria used to arrive at those decisions.
Such opaqueness undermines trust and complicates problem-solving in live applications.
Current intelligent models offer insufficient explainability when what is needed are comprehensible resource allocation decisions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

G. Limited Focus on Small and Medium-Sized Enterprises (SMEs)

However, most of the existing cloud optimization algorithms require large amounts of computing resources. For instance, according to Wang and Yang [2], Zheng et al. [12], Chandrakanth [15], and Zhou et al. [37], there is a prerequisite for large data sets, sufficient computation power, and skilled personnel.

In many cases, small and medium enterprises do not have access to the required resources, which makes it hard to implement complex machine learning pipelines.

Today's research is mostly oriented toward big enterprise environments, paying little attention to the development of lightweight approaches for efficient resource management.

H. Inefficient Handling of Gradual Workload Changes

Current approaches toward the scaling process mainly consider unexpected changes in the workload and their violations of thresholds.

For example, workload changes can be gradual owing to growth in business, seasonal factors, or the behavior of users.

Such studies include those by Roy et al. [35], Shariffdeen et al. [36], and Chanthati [8]; all are oriented to forecasting future loads and give less attention to detecting long-term workload development trends.

As a result, changes in load that take place gradually can go unnoticed until usage of resources reaches critical values.

I. Limited Integration of Trend Analysis with Rule-Based Systems

Rule-based approaches are popular and easy to implement due to their simplicity. Nevertheless, they rely on static thresholds that do not account for workload trends.

Work by Swain et al. [21], Pingulkar et al. [25], and Nagalla et al. [1] confirms the persistent use of threshold-based models in cloud infrastructures.

Even though trend analysis and prediction have been heavily researched separately, only a handful of proposals incorporate both trend-based analysis and lightweight decision engines.

Little work exists on incorporating the analysis of workload trends within rule-based resource management frameworks.

J. Lack of Lightweight Alternatives to Deep Reinforcement Learning

Recent studies have also focused on leveraging deep reinforcement learning (DRL) to optimize resources and improve cloud performance. Works done by researchers including Mao et al. [31], Liu et al. [33], Zhou et al. [37], and Gu et al. [38] highlight the efficiency of applying DRL models.

Nonetheless, the use of DRL involves:

- Long training duration
- High-complexity hardware
- Policy optimization
- Sophisticated deployment steps

As such, their application is not only costly but also complex. There exists a demand for lightweight resource optimization techniques that are not dependent on DRL.

K. Limited Utilization of AWS Monitoring Data for Direct Decision Making

For example, AWS CloudWatch produces an endless flow of data regarding operations performance. Despite the fact that numerous papers work with historical load data to predict further trends, quite a few make use of the available monitoring data and apply it to decision-making.

In most cases, one more layer of machine learning comes before resource allocation is suggested.

Previous research is rather ineffective in terms of direct pattern recognition in cloud monitoring systems.

L. Academia-Industry Gap in Resource Allocation

A considerable number of current cloud resource allocation papers are mainly tested in simulation environments. For instance, Saxena and Singh [11], Bodra and Khairnar [6], and reinforcement learning surveys pay great attention to theoretical analysis of approaches.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The questions of implementation, scalability, easy-to-understand algorithmic solutions, and the ability to operate on real clouds stay out of researchers' primary concerns.
There appears to be a gap between academic approaches to resource allocation and industrial solutions.

VI. FUTURE RESEARCH DIRECTIONS

Despite the usefulness of the proposed framework in demonstrating a successful way of allocating resources in the cloud based on historical records, trend analysis, and decision-making based on predefined rules, there is still ample scope for enhancing the research through further developments. This can be done by employing the latest advancements in technology and achieving better accuracy.

A. Multi-Cloud Resource Allocation

While the existing framework addresses the problem of optimizing the allocation of resources in a single cloud environment, future frameworks may address multi-cloud computing infrastructures that include more than one cloud provider.
Allocation of resources can take into consideration the following aspects among others:

- The cost of services offered by different providers
- Resource availability
- Network latency
- Service reliability

B. Hybrid Cloud and Edge Computing Integration

Today's businesses rely on hybrid cloud and cloud-edge infrastructure. In the future, researchers will be able to modify the framework to allow for effective allocation of resources in these distributed infrastructure scenarios.
Allocation decisions could take into account user proximity, capabilities of the edge devices, available bandwidth, and other considerations. This modification will help to better accommodate IoT, smart city, and real-time analytics use cases.

C. Real-Time Cloud Resource Optimization

The current implementation of the resource management framework analyzes usage patterns based on historical data. The next versions of the framework can use real-time monitoring streams that many modern cloud platforms, such as AWS CloudWatch, provide.
With the ability to process streaming metrics, the framework will be able to detect changing workloads and give timely scaling recommendations.

D. Energy-Efficient Resource Management

Energy is an important issue for today's cloud data center operators. In the future, energy-awareness can be added to the resource allocation framework.
Other potential optimization goals include:

- Power reduction
- Carbon footprint minimization
- Energy efficiency enhancement
- Green cloud computing support

E. Cost and Performance Multi Objective Optimization

While the current proposed methodology concentrates mainly on cost and resource optimization, future researches can explore the concept of multi-objective optimization that includes:

- Resource optimization
- Cost optimization
- Time for response
- Throughput
- Energy consumption

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

- Compliance to SLA

This would allow for more efficient management of resources in cloud computing environments.

F. Automated Cloud Cost Recommendation System

It is anticipated that the next evolution will incorporate the creation of an advanced recommendation system that can automatically suggest ways to optimize costs.

These suggestions can be associated with:

- Right-sizing of the instances
- Utilization of reserved instances
- Use of spot instances
- Storage optimization
- Resource consolidation

G. The Creation of a Fully Automated Cloud Optimization System

One research avenue that can be pursued for many years to come involves creating a full cloud optimization system which incorporates all aspects from monitoring to forecasting and cost analysis into one system.This type of system would enable full automation of cloud resources management in an efficient and cost-effective manner.

VII. CONCLUSION

Cloud computing has evolved into an indispensable element of the modern IT landscape due to its scalability, flexibility, and cost-effectiveness. Nevertheless, proper cloud resource management is one of the key issues since the workload requests are constantly changing and influence both the performance level and overall costs. Incorrect resource allocation could be characterized by over-provisioning that increases the infrastructure costs or under-provisioning which, in turn, reduces application performance levels.

The current research aims at exploring the topic of resource allocation in the cloud and optimizing it to ensure minimum costs. This study included thorough literature research concerning various resource management approaches including traditional algorithms, machine learning, prediction, reinforcement learning, auto-scaling clouds, multi-cloud resources management, hybrid cloud systems, and cloud-edge computing. It has been determined that although such complex approaches as reinforcement learning and machine learning allow making intelligent decisions, they increase computational complexity and require additional training efforts.

Research gaps were observed within the scope of existing cloud resource management schemes. The reactive nature of most legacy systems, which only react to change in workloads, is an area for improvement. Additionally, many existing intelligent systems prioritize prediction accuracy without much emphasis on factors such as simplicity, transparency, interpretability, and deployment. Moreover, most of the existing frameworks utilize large amounts of data and complicated training processes, rendering them unfeasible for implementation under resource-constrained circumstances.

A lightweight and efficient cloud resource allocation scheme was devised to fill the gap described above. The proposed model is based on analysis of the historical usage of resources, analyzing workload patterns, determining trends, and creating rules. In contrast to many machine learning and reinforcement learning models that require training, the proposed approach does not have that requirement. It analyzes historical workload patterns and recognizes trends of increased, decreased, or stable use of resources.

This new framework is an amalgamation of several approaches related to resource management, which includes pattern-based distribution, analysis based on historic data, proactive resource planning, rules-based decision-making, and cost-optimization strategies. As a result, the suggested concept provides a viable compromise between high-level intelligence and easy implementation. Furthermore, the design of this framework is focused on integration capabilities with cloud monitoring services such as AWS CloudWatch.

The main conclusion drawn from the conducted research is that there is important knowledge hidden in historic data that could help make better decisions concerning resource distribution without having to use complicated artificial intelligence methods. Trend analysis and proactive scaling recommendations are aimed at improving resource usage efficiency, cutting redundant provisions, decreasing expenses, and ensuring good application performance.

Overall, the paper introduces a feasible and scalable framework for allocating cloud resources by reconciling the dichotomy between simple thresholds and overly sophisticated machine learning based mechanisms. This framework offers a simple, understandable, and inexpensive alternative to cloud resource optimization that is still easy to implement and maintain.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

As cloud architectures develop in the future, approaches like those offered here will prove to play an increasingly important role in cloud environments.

REFERENCES

[1] S. Nagalla, Y. S. Inturi, and B. S. Inturi, "Adaptive Resource Allocation and Cost Optimization in Cloud Computing," International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 16, no. 2, pp. 310–328, Mar.–Apr. 2025, doi: 10.34218/IJARET_16_02_019.
[2] Y. Wang and X. Yang, "Intelligent Resource Allocation Optimization for Cloud Computing via Machine Learning," Advances in Computer, Signals and Systems, vol. 9, no. 1, pp. 55–63, 2025, doi: 10.23977/acss.2025.090109.
[3] S. Smith, "Optimizing Cost-Aware Cloud Workload Allocation Through Predictive Analytics and Machine Learning–Driven Scheduling Models," International Journal of Cloud Computing, Nov. 2025.
[4] A. Cesarini, "Optimizing Cloud Resource Allocation Using Machine Learning Techniques," Perspective Journal of Computer Science & Systems Biology, vol. 17, no. 2, p. 512, Mar. 2024, doi: 10.37421/0974-7230.2024.17.512.
[5] B. N. Killedar, A. A. Dalvi, and K. M. A. Nazim, "Optimizing Cloud Costs: A Machine Learning–Driven Approach for Efficiency," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 5, no. 10, pp. 31–35, Mar. 2025, doi: 10.48175/IJARSCT-24707.
[6] D. Bodra and S. Khairnar, "Machine Learning-Based Cloud Resource Allocation Algorithms: A Comprehensive Comparative Review," Frontiers in Computer Science, vol. 7, Art. no. 1678976, Oct. 2025, doi: 10.3389/fcomp.2025.1678976.
[7] T. Kamble, S. Deokar, V. S. Wadne, D. P. Gadekar, H. B. Vanjari, and P. Mange, "Predictive Resource Allocation Strategies for Cloud Computing Environments Using Machine Learning," Journal of Electrical Systems, vol. 19, no. 2, pp. 68–77, 2023.
[8] N. S. R. Chanthati, "Predictive Analytics for Cloud Resource Planning and Cost Forecasting," 2025.
[9] Y. Zhang, Y. Gong, J. Xu, B. Liu, J. Huang, and W. Wan, "Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management," 2023.
[10] T. Khan, W. Tian, and R. Buyya, "Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions," arXiv preprint arXiv:2105.05079, May 2021.
[11] D. Saxena and A. K. Singh, "Workload Forecasting and Resource Management Models Based on Machine Learning for Cloud Computing Environments," arXiv preprint arXiv:2106.15112, Jun. 2021.
[12] H. Zheng, K. Xu, M. Zhang, H. Tan, and H. Li, "Efficient Resource Allocation in Cloud Computing Environments Using AI-Driven Predictive Analytics," in Proc. 2nd Int. Conf. on Machine Learning and Automation, 2024, doi: 10.54254/2755-2721/82/2024GLG0055.
[13] S. Deochake, "Cloud Cost Optimization: A Comprehensive Review of Strategies and Case Studies," SSRN Electronic Journal, Aug. 2023.
[14] S. Kayalvili, R. Senthilkumar, S. Yasotha, and R. S. Kamalakannan, "An Optimized Resource Allocation in Cloud Using Prediction-Enabled Reinforcement Learning," Scientific Reports, vol. 15, Art. no. 36088, 2025, doi: 10.1038/s41598-025-19927-2.
[15] L. Chandrakanth, "AI-Driven Dynamic Resource Allocation in Cloud Computing Using Predictive Models and Real-Time Optimization," Journal of Artificial Intelligence, Machine Learning and Data Science, vol. 2, no. 2, pp. 450–456, Jun. 2024, doi: 10.51219/org.doi/JAIMLD/chandrakanth-lekkala/124.
[16] M. Smendowski and P. Nawrocki, "Optimizing Multi-Time Series Forecasting for Enhanced Cloud Resource Utilization Based on Machine Learning," Knowledge-Based Systems, 2024.
[17] R. Yadav and J. S. Yadav, "Leveraging Machine Learning Techniques for Predictive Resource Management in Cloud Environments," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 6, pp. 355–362, Nov.–Dec. 2025.
[18] B. Barua and M. S. Kaiser, "AI-Driven Resource Allocation Framework for Microservices in Hybrid Cloud Platforms," 2024.
[19] A. Rossi, A. Visentin, D. Carraro, S. Prestwich, and K. N. Brown, "Forecasting Workload in Cloud Computing: Towards Uncertainty-Aware Predictions and Transfer Learning," IEEE Transactions on Cloud Computing, 2023.
[20] F. Nwanganga, M. Saebi, G. Madey, and N. Chawla, "A Minimum-Cost Flow Model for Workload Optimization on Cloud Infrastructure," IEEE International Conference on Cloud Computing Technologies and Applications, 2017.
[21] S. R. Swain, A. K. Singh, and C. N. Lee, "Efficient Resource Management in Cloud Environment," arXiv preprint arXiv:2207.12085, 2022.
[22] A. Mukherjee, D. De, and R. Buyya, "Cloud Computing Resource Management," in Resource Management in Distributed Systems, Springer, 2024.
[23] S. H. Anbarkhan, "Optimizing Cloud Resource Allocation with Machine Learning: Strategies for Efficient Computing," Ingénierie des Systèmes d'Information, vol. 30, no. 1, pp. 1–9, 2025.
[24] V. Ramamoorthi, "AI-Driven Cloud Resource Optimization Framework for Real-Time Allocation," Journal of Advanced Computing Systems (JACS), vol. 1, no. 1, pp. 8–15, 2021.
[25] Shriya Pingulkar, Aryaman Tiwary and Shruti Tyagi, "Resource Allocation Techniques in Cloud Computing: A Comprehensive Review," International Journal of Engineering Research & Technology (IJERT), vol. 12, no. 7, pp. 350–356, Jul. 2023.
[26] Rishabh Sinha, "Optimization of Multi-Cloud Workload Placement for Performance and Cost Efficiency"National College of Ireland, Dublin, Ireland, Jan. 2024.
[27] Z. Chen, "Intelligent Resource Management and Optimization in Cloud-Edge Computing" College of Engineering, Mathematics and Physical Sciences, University of Exeter, Sep. 2021.
[28] F.Varghese, "Dynamic Resource Allocation in Multi-Cloud Environments Using Reinforcement Learning," MSc Research Project, Masters in Cloud Computing, School of Computing, National College of Ireland, 2024.
[29] S. Malik, M. Tahir, M. Sardaraz, and A. Alourani, "A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques," Applied Sciences, vol. 12, no. 4, p. 2160, Feb. 2022, doi: 10.3390/app12042160.
[30] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, Oct.–Dec. 2021, doi: 10.1016/j.ijforecast.2021.03.012.
[31] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource Management with Deep Reinforcement Learning," in Proceedings of the 15th ACM Workshop on Hot Topics in Networks (HotNets-XV), Atlanta, GA, USA, Nov. 2016, pp. 50–56, doi: 10.1145/3005745.3005750.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[32] Y. Ye, X. Ren, J. Wang, L. Xu, W. Huang, and W. Tian, "A New Approach for Resource Scheduling with Deep Reinforcement Learning," in 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), Singapore, Dec. 2018, pp. 122–129, doi: 10.1109/PADSW.2018.8645040.

[33] N. Liu, Z. Li, J. Xu, Z. Xu, S. Lin, Q. Qiu, J. Tang, and Y. Wang, "A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning," in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, Jun. 2017, pp. 372–382, doi: 10.1109/ICDCS.2017.42.

[34] Y. Garí, D. A. Monge, E. Pacini, C. Mateos, and C. G. Garino, "Reinforcement Learning-Based Application Autoscaling in the Cloud: A Survey," Computer Science Review, vol. 37, Art. no. 100273, Aug. 2020, doi: 10.1016/j.cosrev.2020.100273.

[35] N. Roy, A. Dubey, and A. Gokhale, "Efficient Autoscaling in the Cloud Using Predictive Models for Workload Forecasting," in 2011 IEEE 4th International Conference on Cloud Computing (CLOUD), Washington, DC, USA, Jul. 2011, pp. 500–507, doi: 10.1109/CLOUD.2011.79.

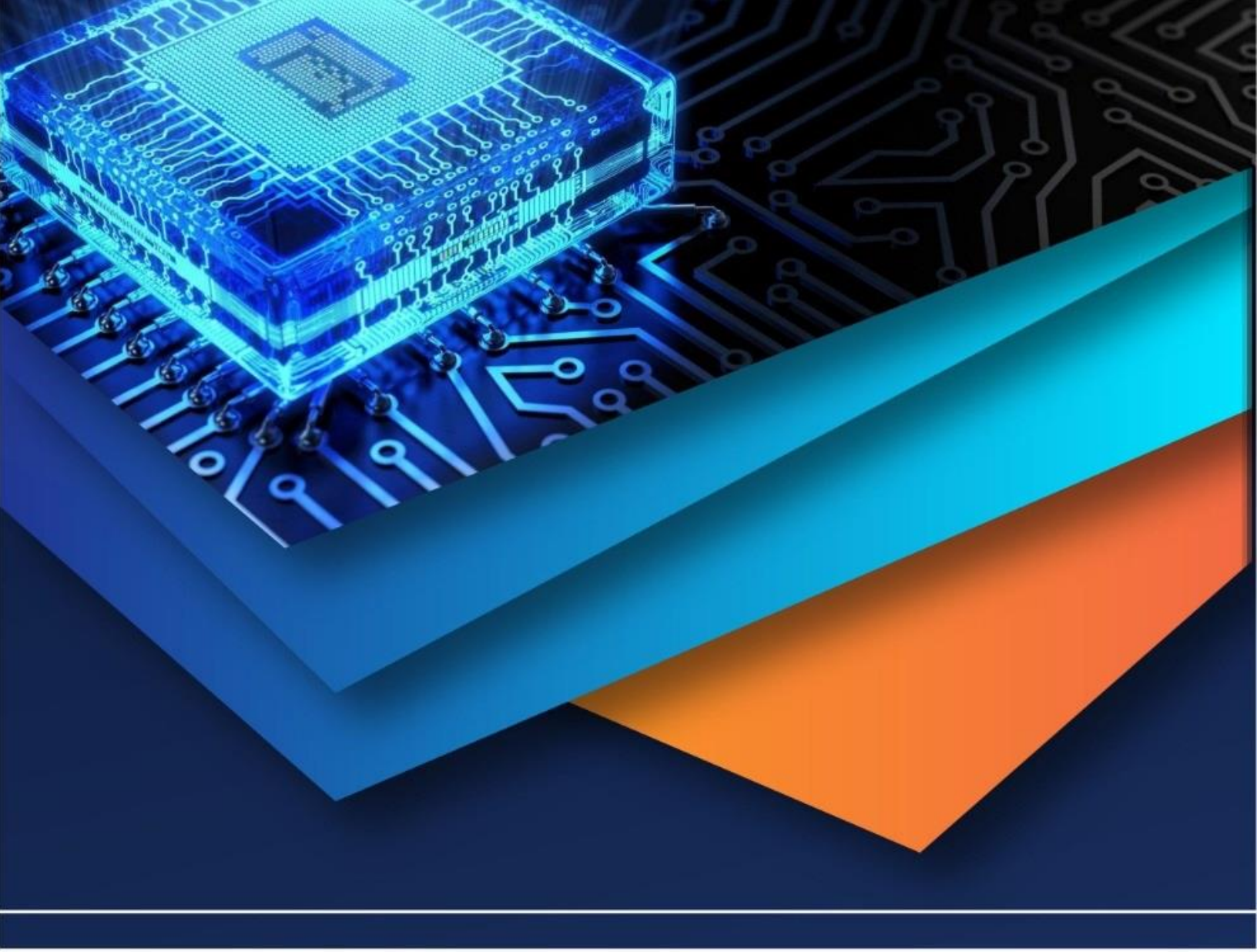
[36] R. S. Shariffdeen, D. T. S. P. Munasinghe, H. S. Bhathiya, U. K. J. U. Bandara, and H. M. N. Dilum Bandara, "Adaptive Workload Prediction for Proactive Auto Scaling in PaaS Systems," in 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Luxembourg, Dec. 2016, pp. 336–343, doi: 10.1109/CloudCom.2016.0062.

[37] G. Zhou, W. Tian, R. Buyya, R. Xue, and L. Song, "Deep Reinforcement Learning-Based Methods for Resource Scheduling in Cloud Computing: A Review and Future Directions," Artificial Intelligence Review, vol. 57, Art. no. 124, 2024, doi: 10.1007/s10462-024-10756-9.

[38] Y. Gu, Z. Liu, S. Dai, C. Liu, Y. Wang, S. Wang, G. Theodoropoulos, and L. Cheng, "Deep Reinforcement Learning for Job Scheduling and Resource Management in Cloud Computing: An Algorithm-Level Review," arXiv preprint arXiv:2501.01007, Jan. 2025.

[39] S. Alharthi, A. Alshamsi, A. Alseiari, and A. Alwarafy, "Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions," Sensors, vol. 24, no. 17, p. 5551, Aug. 2024, doi: 10.3390/s24175551.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)