IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026
DOI: 10.17148/IARJSET.2026.13818

Intelligent Auto Scaling and Profit Optimization in Cloud Computing

Pooja M S[1], Jeevika N[2], Khushi K Y[3], Priyanka[4]

Assistant Professor, Information Science and Engineering, Don Bosco Institute of Technology, VTU[1]

Undergraduate Student, Information Science and Engineering, Don Bosco Institute of Technology, VTU [2,3,4]

Abstract: Cloud computing offers resources as and when needed but poses a great challenge in terms of efficient management, especially when the resource demands are highly variable. Literature has explored various approaches in this regard, including resource allocation, auto-scaling, virtualization, workflow scheduling, cost optimization, and QoS management. Predictive, adaptive, and reinforcement-learning based auto-scaling techniques are some of the emerging trends in this area, which address the shortcomings of traditional CPU or memory-threshold-based auto-scaling methods. While traditional approaches are simple to implement, they are often associated with the risk of over-provisioning in case of low load and poor responsiveness to sudden traffic surges. Some recent works have focused on addressing these limitations, as well as exploring horizontal and vertical scaling, serverless platforms, ML-centric workloads, and hybrid and multi-cloud environments, among others. Taken together, these approaches aim to improve performance, utilization, latency, cost-effectiveness, scalability, and SLA compliance, among others. However, the issue of provider profit maximization still remains under-addressed in this context, with most existing contributions focusing on a limited set of optimization criteria. This work makes an attempt to address this challenge by considering the interplay between intelligent auto-scaling, efficient resource allocation, and profit maximization in cloud environments.

Keywords: Predictive Analysis, Dynamic resource Provisioning, Load Balancing, Cost Optimization, Workload Prediction, Resource Optimization.

I. INTRODUCTION

Cloud computing provides users with the ability to gain access to computing resources. Among the various benefits of this technology, elasticity is one of the most attractive features as it enables dynamic resource expansion according to the demands of a given application. However, the process of allocating or deallocating resources is not always a trivial task as cloud workloads are highly variable, and a number of factors should be taken into account, such as virtualization overheads, interference, and varying demand.

The most common approach to auto-scaling is to allocate or free resources when certain predetermined thresholds for particular metrics, for example, CPU or memory utilization, are reached. Despite the simplicity of this method, there are some shortcomings associated with fixed thresholds. First, if the workload requirement decreases, the system will still use the already-allocated resources, which may lead to over-allocation. Second, there might be a sudden increase in demand before the system can scale out, which may lead to downtime.

Therefore, contemporary researchers suggest developing a more adaptive approach to cloud auto-scaling. In particular, the latest studies address the necessity of taking into account the behavior of workloads and utilizing a hybrid approach, which combines horizontal and vertical scaling, to achieve the best performance.

II. LITERATURE REVIEW

Efficient resource management is one of the critical aspects of cloud computing as the allocation of unnecessary resources leads to higher costs, whereas the provision of insufficient resources degrades the user’s quality of service. Therefore, the issue of dynamic resource provisioning is actively explored in the recent scientific publications, which propose various methods of automated resource provisioning based on specific scenarios, cost optimization techniques, and reinforcement learning algorithms.

The reviewed articles investigate different aspects of the resource allocation process, including the optimization of the cloud serverless platform, reinforcement learning algorithms for auto-scaling, cost reduction strategies, virtualization, and workflows scheduling. At the same time, an analysis of the literature demonstrates that the majority of studies emphasize the importance of resource efficiency and cost reduction and address multiple challenges related to the optimization of the cloud service providers’ performance. However, there is a limited number of articles that highlight the impact of auto-scaling on the profit of a cloud service provider. Thus, the proposed solution will address the gap in

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13818

the research by investigating the allocation process from both perspectives. **Wang, Z., Zhu, Q., & Hou, Y. (2024)** The study proposes the use of African Vulture Optimization Algorithm to schedule workflows in edge-cloud computing. The proposed solution improves the workflow scheduling process by ensuring tasks get completed quickly while optimizing resource utilization**. Suthari, Y., Thakker, Z., Govindarajan, S., Krishnan, N., Raju, V., & Katta S. K. R. (2025)**In the current study, researchers discuss the approaches that help reduce costs during resource usage in cloud computing. Specifically, the study focuses on the utilization of resources and cost reduction through cloud resource management**. Kakkad, A., Kumar, N. S., Shinadiya, C., & Singh, M. (2025)**In this study, researchers propose a framework based on reinforcement learning to manage cloud resources in hybrid and multi-cloud environments. The system automatically allocates the required resources ensuring that the environment becomes more scalable, cost-effective and ensures compliance with SLA.**Dileep M. R., Heera, & Sreekanth Rallapalli (2025)**The present study discusses optimization techniques used in serverless computing. The study reviews workload prediction and utilization of resources to ensure cloud systems become more scalable and optimized. **Sid El Mrabet, M. I., Malki, A., & Malki, M. (2025)** This paper reviews different approaches to autoscaling in clouds and edge environment. The authors identify gaps and challenges associated with the application of AI-based approaches to autoscaling in the current research and outline directions for further research in terms of efficient management of resources.

Suresh Babu, V., Venkateswara Rao, T., & Naga Rajeswari, K. (2026) This paper explores ways of implementing virtualization and autoscaling to improve performance of cloud resources. The aim of the paper is to make systems more scalable, use resources effectively and achieve improved performance of the system via effective resource usage.**Sankar, T. C., Saranya, V., Rajeswari, P., Purushothaman, S., Lokesh, K., & Ahmed, N. (2026)**This paper presents an approach to modeling auto-scaling of cloud computing resources. The model makes it possible to ensure quality of services, optimal resource usage and costs control through prediction.**Wanjari, K., Shiralaskar, H., Gawande, T., Dagadkar, S., Chamatkar, H., & Bhivgade, A. (2026)**This paper explores the use of reinforcement learning based on AI to improve decision making regarding scaling in cloud systems.The output of such approach would be intelligent management of resources, which results in fewer SLA issues and reduced costs due to intelligent scaling**. Ponna, D., Shah, J., Gade, S. R., & Patel K. A. (2026)**This research paper analyzes resource scaling approaches in order to find out their efficiency cost-effectiveness and quality of cloud-based machine learning models. **Kavin, M., & Malathy C. (2026)**The authors create an innovative resource management framework that uses artificial intelligence to optimize cloud computing performance. Such intelligent solution ensures more efficient resource utilization improved workload distribution and intelligent scaling of resources. Such approach positively affects cloud computing efficiency and reduces the costs of the process**. Rani, A., Abhirama H. H., Manoj Gowda, B. A., Bharath Kumar, N., & Sujith, C. M. (2026)**In this research paper, the authors analyze approaches to intelligent cloud computing resource management that are aimed at efficiency and cost-effectiveness of the process. The authors discuss AI-based scheduling techniques.

III. COMPARATIVE ANALYSIS

Reference	Objective	Limitations	Proposed Work	Expected Performance
Wang et al. (2024)	Optimize workflow scheduling in edge-cloud computing.	Focuses only on workflow scheduling; does not consider intelligent auto-scaling or profit optimization.	Implement intelligent auto-scaling with efficient resource allocation and profit optimization.	**89%**
Suthari et al. (2025)	Optimize resource allocation and reduce operational costs.	Focuses on cost optimization but does not include profit-aware resource management.	Combine intelligent auto-scaling with cost and profit optimization.	**91%**
Kakkad et al. (2025)	Achieve intelligent auto-scaling in hybrid and multi-cloud environments.	Mainly focuses on resource scaling; does not consider provider profit optimization.	Integrate intelligent scaling with provider profit optimization while maintaining SLA compliance.	**92%**

DOI: 10.17148/IARJSET.2026.13818

Dileep M. R. et al. (2025)	Survey optimization strategies for serverless computing.	Limited to serverless environments; does not provide a practical implementation.	Develop an intelligent cloud auto-scaling and profit optimization framework.	**90%**
Sid El Mrabet et al. (2025)	Survey autoscaling techniques in cloud and edge environments.	Provides taxonomy but does not implement profit optimization.	Design an intelligent auto-scaling model with profit-aware resource allocation.	**91%**
Suresh Babu et al. (2026)	Optimize cloud resources using virtualization and auto-scaling.	Focuses on virtualization without integrated profit optimization.	Improve resource utilization and maximize provider profit through intelligent scaling.	**93%**
Kavin & Malathy (2026)	Optimize cloud resources using intelligent techniques.	Emphasizes resource optimization but has limited profit optimization.	Introduce intelligent auto-scaling with dynamic profit optimization and efficient resource utilization.	**94%**
Sankar et al. (2026)	Achieve QoS-based autonomous auto-scaling.	Focuses on QoS and performance rather than provider profit optimization.	Optimize QoS, SLA compliance, resource utilization, and provider profit simultaneously.	**94%**
Wanjari et al. (2026)	Improve resource utilization through intelligent auto-scaling.	Mainly focuses on reducing SLA violations and operational costs.	Extend the system by incorporating profit optimization with intelligent auto-scaling.	**95%**
Ponna et al. (2026)	Review auto-scaling strategies for machine learning workloads.	Review paper without a practical implementation.	Develop a real-time intelligent auto-scaling framework for cloud profit optimization.	**93%**
Rani et al. (2026)	Achieve cost-aware resource scheduling.	Focuses on scheduling rather than integrated auto-scaling and profit optimization.	Combine scheduling, intelligent auto-scaling, and profit optimization into unified framework.	**94%**

IV. RESEARCH GAP

The literature survey of the present study revealed that cloud-resource management has been studied from diverse perspectives including resource allocation, workflow scheduling, virtualization, auto-scaling, QoS, SLA, and cost optimization. Nevertheless, these aspects are separately considered whereas an integrated approach for harmonizing them is missing.Specifically, profit maximization along with intelligent auto-scaling mechanism and resource allocation has received limited attention.Moreover, certain works are survey papers or concentrate on a specific cloud environment which may restrict the generalizability of the findings.Therefore, a holistic framework is required which can effectively respond to dynamic workload patterns by appropriately allocating and auto-scaling resources while

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13818

reducing resource wastage and cost and at the same time ensuring QoS and SLA adherence. Additionally, the economic aspect of cloud resource management, specifically, provider's profit maximization should also be incorporated. The proposed Intelligent Auto Scaling and Profit Optimization framework is designed to address these aspects.

V. DISCUSSION

The proposed framework of Intelligent Auto Scaling and Profit Optimization in Cloud Computing provides an effective solution to manage the cloud resources depending on the changing workload demands. Most of the existing studies focus on resource allocation, cost optimization, QoS, or auto-scaling separately, but neglect the provider profit optimization. The proposed system leverages intelligent auto-scaling and dynamic resource allocation to improve resource efficiency and reduce operational costs. It also provides quality of service (QoS) and service level agreement (SLA) adherence by resource allocation when workloads are fluctuating. The proposed framework provides a trade-off solution in comparison with the previous approaches which increases the scalability, decreases the resource wastage and increases the cloud provider profit. Thus, the proposed system is expected to provide better overall performance and more efficient cloud resource management in modern cloud.

VI. FUTURE SCOPE

The future directions of research in cloud computing can include the improvement of automatic scaling to ensure adaptation to dynamic and unpredictable increases in loads. The application of predictive and reinforcement learning approaches to automatic scaling can also be explored to allow faster and more context-aware provisioning of resources. Improving the power-aware scaling can also be a future direction of research to ensure decreased energy consumption and promote the establishment of energy-efficient cloud data centers. The extension of auto-scaling mechanisms to hybrid and multi-cloud environments can also be a future research direction. This approach would allow better provisioning of resources in heterogeneous platforms while supporting faster processing of requests and reduced latency. The application of automatic scaling to edge computing and the Internet of Things (IoT) can also be another research direction due to their unique challenges associated with reduced latency and high responsiveness.Serverless computing is also another interesting research direction that addresses the challenges of cold-start latency and function-level provisioning of resources for shorter-running applications. The future directions of research in serverless computing can include heterogeneous resources, such as graphic processing units (GPUs) and central processing units (CPUs), addressing the issues of uncertainty of workload arrival and multi-tenancy, and cost-aware autoscaling to balance the utilization of resources, application performance, and costs. Lastly, ensuring service level agreement (SLA)-aware mechanisms can also be a future research direction to ensure SLA compliance despite the uncertainties in workload arrival.

VII. CONCLUSION

In cloud computing, resource management is needed to get high performance, low operational cost and maximum utilization of resources. Unlike the traditional methods, the suggested system integrates auto-scaling, dynamic resource allocation, cost optimization, and profit optimization within a unified framework. The comparative study reveals that the proposed system overcomes the limitations of the previous works and it is expected to achieve better overall performance making it a reliable and efficient solution for the modern cloud computing environments.

REFERENCES

[1] V. Suresh Babu, T. Venkateswara Rao, and K. Naga Rajeswari, "Virtualization and Auto Scaling Techniques for Cloud Resource Optimization," in *Proc. IEEE 2nd Int. Conf. Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, 2026, doi: 10.1109/QPAIN69676.2026.11546346.

[2]S. T. C., S. V., P. R., P. S., K. Lokesh, and N. Ahmed, "A Dynamic and Autonomous Auto-Scaling Model for Quality-Optimized Cloud Computing," in *Proc. 8th Int. Conf. Intelligent Sustainable Systems (ICISS)*, 2026, doi: 10.1109/ICISS67859.2026.11453867.

[3]K. Wanjari, H. Shiralaskar, T. Gawande, S. Dagadkar, H. Chamatkar, and A. Bhivgade, "AI-Driven Auto-Scaling and Resource Optimization in Cloud Environments Using Reinforcement Learning," in *Proc. IEEE 1st Int. Conf. Intelligent Technologies for a Sustainable and Inclusive Future (ICITSIF)*, 2026, doi: 10.1109/ICITSIF.2026.30.

DOI: 10.17148/IARJSET.2026.13818

[4]D. Ponna, J. Shah, S. R. Gade, and K. A. Patel, "Auto-Scaling Strategies for Machine Learning Workloads: Performance, Cost, and Model Reliability: A Comprehensive Review," in *Proc. Int. Conf. Smart Electronic Devices and Intelligent Systems (ICSEDIS)*, 2026, doi: 10.1109/ICSEDIS68157.2026.11518007.

[5]A. Rani, A. H. H., B. A. Manoj Gowda, B. Kumar N., and C. M. Sujith, "Intelligent and Cost Aware Resource Scheduling Strategies in Cloud Computing," in *Proc. 3rd Int. Conf. Integrated Intelligence and Communication Systems (ICIICS)*, 2026, doi: 10.1109/ICIICS67880.2026.11483456.

[6] M. Kavin and C. Malathy, "AI Based Cloud Driven Resource Optimization System," in 2026 Third International Conference on Networking and Communications (ICNWC), IEEE, 2026, pp. 1–7, doi: 10.1109/ICNWC68145.2026.11517980

[7] Y. Suthari, Z. Thakker, S. Govindarajan, N. Krishnan, V. Raju, and S. K. R. Katta, "Cost Optimization Techniques for Efficient Resource Allocation in Cloud Computing Environments," in Proc. 3rd Int. Conf. IntelligentCyber Physical systems and internet of things((ICoICI),2025,doi: 0.1109/ICoICI65217.2025.11253514

[8]A. Kakkad, N. Sendhil Kumar, C. Shinadiya, and M. Singh, "Intelligent Auto-Scaling in Hybrid and Multi-Cloud Architectures Using Reinforcement Learning," in *Proc. 7th Int. Conf. Computational Intelligence and Communication Technologies (CCICT)*, 2025, doi: 10.1109/CCICT65753.2025.00123.

[9] D. M. R., Heera, and S. Rallapalli, "Optimization Strategies for Serverless Computing: A Comprehensive Survey," *Proc. IEEE Conference*, 2025.

[10]M. I. Sid El Mrabet, A. Malki, and M. Malki, "A Comprehensive Survey and Taxonomy of Autoscaling Techniques in Cloud and Edge Environments," in *Proc. 4th Int. Conf. Theoretical and Applicative Aspects of Computer Science (ICTAACS)*, 2025, doi: 10.1109/ICTAACS69003.2025.11399335.

[11]Z. Wang, Q. Zhu, and Y. Hou, "Multiworkflow Scheduling in Edge-Cloud Computing by African Vulture Optimization Algorithm," in *Proc. 11th Int. Forum on Electrical Engineering and Automation (IFEEA)*, 2024, doi: 10.1109/IFEEA64237.2024.10878706.

