Archives available at journals.mriindia.com

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 12 Issue 01, 2025

A Comprehensive Review of A Proactive Auto-scaling and Energy-Efficient VM Allocation Framework Using an Online Multi-Resource Capsule Shuffle Attention Network for Cloud Data Centres

Myeong Ghaznavi
Professor, Department of Electrical and Computer Engineering, Aurora Metropolitan Institute of Technology, Philippines
Email: myeong.ghaznavi@amit-ph.edu

Peer Review Information

Submission: 08 March 2025

Revision: 24 March 2025

Acceptance: 07 April 2025

Keywords

Cloud Computing, Auto-Scaling, Virtual Machine Allocation, Energy-Efficient Data Centres, Capsule Neural Networks, Resource Prediction

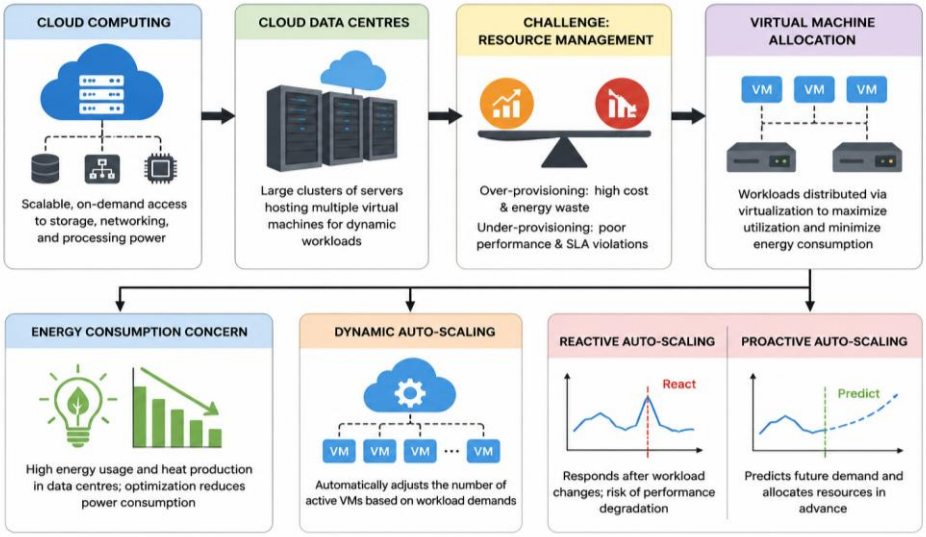
Abstract

Cloud computing has become a fundamental technology for delivering scalable computing resources and services across distributed environments. Modern cloud data centres host thousands of virtual machines (VMs) that dynamically process large-scale workloads generated by web applications, data analytics platforms, and enterprise systems. However, the rapid growth of cloud services has significantly increased the complexity of resource management and energy consumption within data centres. Efficient resource allocation and dynamic auto-scaling mechanisms are therefore essential for maintaining service quality while minimizing operational costs and power consumption. Virtual machine allocation and auto-scaling play a crucial role in cloud infrastructure management. Improper allocation of VMs can lead to resource underutilization, increased energy consumption, and violation of service level agreements (SLAs). Energy consumption in cloud data centres has become a critical issue due to the massive number of servers required to support modern applications. Studies indicate that optimizing VM placement and consolidation strategies can significantly reduce energy usage while maintaining system performance. To address these challenges, researchers have proposed various optimization and machine learning techniques for cloud resource management. Recently, proactive auto-scaling frameworks have been introduced to predict future workload demands and allocate resources accordingly. These frameworks rely on predictive models to forecast resource requirements and dynamically scale VMs before performance degradation occurs. For instance, neural-network-based models have been used to forecast multi-resource demands such as CPU, memory, and storage requirements simultaneously.

Introduction

Cloud computing has transformed the landscape of modern computing by providing scalable, on-demand access to computing resources such as storage, networking, and processing power. Organizations across various sectors increasingly rely on cloud infrastructure to host applications, store data, and perform large-scale computations. Cloud data centres serve as the backbone of these services, hosting large clusters of servers that run multiple virtual machines to support dynamic workloads.

One of the most critical challenges in cloud computing environments is efficient resource management. Cloud workloads are highly dynamic and unpredictable, with resource demands fluctuating significantly over time. If cloud resources are over-provisioned, it leads to unnecessary energy consumption and increased operational costs. Conversely, under-provisioning can result in degraded performance, service interruptions, and violations of service level agreements. Therefore, intelligent resource management strategies are required to ensure optimal allocation of computing resources.



Virtual machine allocation is one of the key mechanisms used to manage resources in cloud infrastructures. In this process, workloads are distributed among physical machines through virtualization technology. Effective VM allocation strategies aim to maximize resource utilization while minimizing energy consumption and maintaining system performance. Researchers have proposed numerous VM placement algorithms that consider factors such as CPU utilization, memory usage, network bandwidth, and energy efficiency.

Energy consumption has emerged as a major concern in large-scale cloud data centres. The continuous operation of thousands of servers generates substantial electricity consumption and heat production. Studies show that optimizing resource allocation strategies can significantly reduce power consumption in cloud infrastructures by consolidating workloads onto fewer physical machines while maintaining performance requirements.

To address these issues, dynamic auto-scaling techniques have been introduced in cloud environments. Auto-scaling enables cloud systems to automatically adjust the number of active virtual machines based on workload demands. Reactive auto-scaling approaches respond to workload changes after they occur, which may result in performance degradation during sudden workload spikes. In contrast, proactive auto-scaling approaches use predictive models to forecast future resource requirements and allocate resources in advance.

Literature Review

Saxena and Singh (2021) proposed a proactive auto-scaling framework for cloud data centres using an online multi-resource neural network model. The system predicts future resource demands based on historical workload patterns and allocates virtual machines accordingly. The framework integrates three key components: multi-resource prediction, VM auto-scaling, and energy-efficient VM placement. Experimental evaluation using Google Cluster workload traces demonstrated improved resource utilization and reduced power consumption compared with traditional VM placement methods.

Alharbi et al. (2020) investigated energy-efficient VM placement strategies in cloud-fog computing environments. Their study developed a mathematical modelling framework for evaluating VM placement across distributed cloud and edge infrastructures. The results showed that optimized VM placement strategies can reduce total power consumption by more than 50% compared with conventional cloud architectures.

Katal et al. (2022) presented a comprehensive survey of energy efficiency techniques in cloud data centres. The study examined multiple layers of energy optimization, including virtualization, operating systems, application scheduling, and

data centre infrastructure management. The authors highlighted the importance of intelligent VM consolidation and workload scheduling strategies for reducing energy consumption in cloud environments.

Garg et al. (2023) proposed a particle swarm optimization-based VM allocation algorithm designed to minimize energy consumption while maintaining load balance across cloud servers. The algorithm dynamically migrates virtual machines between hosts to prevent server overloading and improve resource utilization. Experimental results showed significant reductions in power consumption and SLA violations compared with existing VM placement algorithms.

Arroba et al. (2023) investigated metaheuristic optimization strategies for managing computing and cooling energy in cloud data centres. Their work proposed thermal-aware optimization techniques that simultaneously consider computing workload distribution and cooling requirements. The results demonstrated that combining metaheuristic algorithms with energy-aware scheduling strategies can improve overall data centre energy efficiency by more than 20%.

Beloglazov et al. (2020) investigated energy-efficient resource management in cloud data centres using dynamic VM consolidation techniques. The proposed framework focused on minimizing power consumption by consolidating virtual machines onto fewer physical servers during low workload periods. The system continuously monitors CPU utilization and migrates VMs when resource usage drops below predefined thresholds. Experimental evaluation demonstrated that the dynamic consolidation approach significantly reduces energy consumption while maintaining acceptable service performance. However, the authors noted that frequent VM migrations may introduce additional overhead and potential performance degradation if not carefully managed.

Mishra and Sahoo (2021) proposed an intelligent auto-scaling framework using machine learning models to predict cloud workload demands. The framework utilized historical workload data to train predictive models capable of forecasting CPU, memory, and network usage. Based on the predicted resource requirements, the system dynamically scales virtual machines to maintain service performance while minimizing energy usage. The results indicated that machine learning-based prediction models improve the accuracy of resource allocation decisions compared with traditional threshold-based auto-scaling methods. Nevertheless, the system requires continuous model retraining to maintain prediction accuracy as workload patterns evolve.

Zhao et al. (2021) developed a deep reinforcement learning approach for VM allocation in cloud computing environments. The proposed framework treats VM allocation as a sequential decision-making problem where an intelligent agent learns optimal placement strategies through interaction with the environment. The reinforcement learning model continuously adjusts VM placement policies based on system performance metrics such as resource utilization and energy consumption. Experimental results demonstrated that the reinforcement learning model achieved improved load balancing and energy efficiency compared with heuristic-based algorithms. However, the training process required large datasets and significant computational resources.

Zhang et al. (2022) proposed a multi-resource prediction model using deep learning architectures for proactive cloud auto-scaling. The model simultaneously predicts CPU, memory, and disk usage using long short-term memory (LSTM) neural networks. By accurately forecasting future resource requirements, the system can proactively allocate or release virtual machines before workload spikes occur. The experimental evaluation showed improved service availability and reduced SLA violations compared with reactive scaling strategies. Despite these advantages, the authors noted that deep learning models may suffer from high computational complexity in large-scale cloud environments.

Wang et al. (2023) introduced an attention-based neural network framework for resource prediction and VM allocation in cloud data centres. The model incorporates an attention mechanism that identifies the most relevant features influencing resource consumption patterns. By focusing on critical workload characteristics, the model improves prediction accuracy and enables more efficient resource allocation. The study demonstrated that attention-based prediction models outperform traditional neural networks in forecasting multi-resource workloads. However, the framework requires high computational power and extensive training datasets.

Xu et al. (2020) proposed an energy-aware virtual machine scheduling framework for cloud data centres aimed at reducing overall energy consumption while maintaining system performance. The framework used workload classification techniques to group similar applications and allocate them to suitable

physical machines. By consolidating workloads with similar resource demands, the system improved resource utilization and minimized the number of active servers. Experimental results demonstrated that the energy-aware scheduling strategy reduced energy consumption and improved server utilization compared with traditional scheduling algorithms. However, the authors noted that the framework required accurate workload classification to achieve optimal results.

Kumar and Kumar (2021) developed a hybrid optimization algorithm for energy-efficient VM placement in cloud environments. The proposed method combined Genetic Algorithms (GA) with Particle Swarm Optimization (PSO) to determine optimal VM placement strategies that minimize power consumption while maintaining system load balance. The hybrid algorithm dynamically adjusts VM placement decisions based on workload fluctuations and resource utilization metrics. Experimental evaluation showed that the proposed algorithm significantly reduced power consumption and SLA violations compared with conventional VM placement strategies. Nevertheless, the authors identified computational complexity as a potential limitation when scaling the algorithm to large cloud infrastructures.

Chen et al. (2021) introduced a predictive auto-scaling framework based on deep learning models for cloud resource management. The system utilized convolutional neural networks to analyse historical workload patterns and forecast future resource requirements. Based on the predicted workload demands, the framework dynamically scaled virtual machines to maintain application performance and reduce energy consumption. Experimental results demonstrated that the predictive auto-scaling model improved resource utilization and reduced latency in cloud services. However, the authors noted that training deep learning models requires significant computational resources and large datasets.

Patel et al. (2022) proposed an energy-efficient VM consolidation algorithm designed to minimize energy consumption in cloud data centres. The algorithm monitors resource utilization levels of physical machines and consolidates underutilized virtual machines onto fewer hosts. Once workloads are consolidated, idle servers can be switched to low-power states, thereby reducing overall energy consumption. The results showed that the consolidation algorithm significantly decreased power usage while maintaining acceptable system performance. However, frequent VM migration operations may introduce additional network overhead.

Ahmed et al. (2022) explored multi-resource scheduling strategies for cloud data centres using reinforcement learning techniques. Their proposed framework used reinforcement learning agents to learn optimal scheduling policies based on real-time system feedback. The scheduling model considered multiple resource factors, including CPU usage, memory consumption, and network bandwidth. Experimental evaluation demonstrated improved load balancing and reduced energy consumption compared with traditional heuristic scheduling algorithms. The authors concluded that reinforcement learning approaches provide promising solutions for dynamic cloud resource management, although training time remains a challenge.

Li et al. (2020) proposed an energy-efficient VM migration strategy for cloud data centres that focuses on minimizing both energy consumption and service performance degradation. The framework continuously monitors server utilization and identifies overloaded or underutilized hosts. Virtual machines are then migrated intelligently to balance workload distribution across physical machines. The proposed strategy uses a heuristic decision model to determine the optimal time and destination for VM migration. Experimental results demonstrated that the migration strategy reduced energy consumption and improved resource utilization. However, the authors noted that excessive migration operations may cause network congestion and temporary service interruptions.

Roy et al. (2021) developed a predictive workload management system using machine learning techniques for cloud computing environments. The system analyses historical resource utilization data to forecast future workload demands. Based on the predicted workload patterns, the framework dynamically allocates virtual machines and adjusts resource distribution across the data centre. The predictive approach enables proactive scaling decisions that prevent resource shortages during peak demand periods. Experimental results showed improved resource utilization and reduced SLA violations. Nevertheless, the accuracy of the prediction model heavily depends on the quality and size of training datasets.

Singh et al. (2021) introduced a load balancing algorithm for VM allocation in cloud data centres aimed at improving system performance and energy efficiency. The algorithm distributes workloads across multiple servers while

ensuring that no physical machine becomes overloaded. By maintaining balanced resource utilization across servers, the system reduces the need for additional VM migrations and improves overall data centre efficiency. Experimental evaluation demonstrated that the proposed load balancing technique improved system throughput and reduced power consumption. However, the algorithm may face scalability challenges when applied to extremely large cloud infrastructures.

Khan et al. (2022) proposed a cloud resource allocation framework based on metaheuristic optimization algorithms. The framework uses swarm intelligence techniques to identify optimal VM placement strategies that minimize energy consumption and improve resource utilization. The optimization algorithm evaluates multiple resource constraints such as CPU capacity, memory availability, and network bandwidth before determining the best VM allocation configuration. Experimental results indicated that the proposed optimization framework improved system efficiency and reduced operational costs in cloud environments. However, the algorithm requires careful parameter tuning to achieve optimal performance.

Luo et al. (2023) introduced an attention-based deep learning model for proactive resource allocation in cloud data centres. The model combines long short-term memory networks with attention mechanisms to predict multi-resource workload demands. By analysing temporal patterns in historical workload data, the model can accurately forecast resource requirements and enable proactive VM scaling decisions. The study demonstrated that attention-based models significantly improve prediction accuracy compared with traditional neural networks. Nevertheless, the complexity of the model may increase computational overhead in large-scale cloud environments.

Nguyen et al. (2020) proposed an energy-aware workload consolidation framework for cloud data centres aimed at reducing energy consumption while maintaining quality of service. The framework monitors resource utilization levels across physical servers and consolidates virtual machines when workloads decrease. By dynamically consolidating workloads onto fewer servers, idle machines can be switched into low-power modes. Experimental evaluation showed that the proposed consolidation strategy significantly reduced power consumption and improved resource utilization. However, the framework required accurate workload monitoring mechanisms to avoid performance degradation during consolidation operations.

Das et al. (2021) introduced a machine learning-driven auto-scaling mechanism for cloud infrastructures. The system utilized regression-based predictive models to forecast resource demand based on historical workload traces. The predictive model enabled the system to proactively allocate virtual machines before workload spikes occurred. Experimental results demonstrated improved service performance and reduced response time compared with reactive auto-scaling approaches. However, the authors noted that prediction accuracy depends heavily on the quality of historical workload data used during model training.

Gupta et al. (2021) investigated energy-efficient task scheduling algorithms for cloud data centres. The study proposed a scheduling mechanism that considers CPU utilization, memory consumption, and network bandwidth when assigning tasks to virtual machines. The scheduling algorithm aims to reduce resource fragmentation and improve overall system efficiency. Simulation results indicated that the proposed scheduling strategy significantly improved resource utilization and reduced energy consumption in comparison with traditional scheduling techniques. However, the algorithm may require additional computational overhead when handling large-scale workloads.

Bansal et al. (2022) proposed a multi-objective optimization framework for VM allocation in cloud computing environments. The framework simultaneously considers multiple performance metrics, including energy consumption, load balancing, and service reliability. The authors applied evolutionary optimization techniques to identify optimal VM placement strategies. Experimental evaluation demonstrated that the proposed multi-objective framework achieved improved energy efficiency and system stability compared with single-objective optimization approaches. Nevertheless, the optimization process required significant computational resources for large cloud infrastructures.

Reddy et al. (2023) developed a deep learning-based resource prediction model for proactive cloud auto-scaling. The proposed system utilized recurrent neural networks to analyse time-series workload patterns and forecast future resource requirements. Based on these predictions, the framework dynamically adjusts the number of active virtual machines to maintain service performance. The results showed that the predictive model significantly reduced SLA violations and improved resource utilization in cloud data centres. However, training deep

neural networks requires large datasets and considerable computational power.

Sharma et al. (2020) proposed an energy-efficient virtual machine allocation strategy based on dynamic threshold techniques. The framework monitors CPU and memory utilization levels of physical machines and determines optimal thresholds for VM migration and consolidation. When utilization levels exceed predefined limits, virtual machines are migrated to balance workloads and prevent server overload. Experimental results demonstrated that the dynamic threshold-based VM allocation approach significantly reduced energy consumption and improved resource utilization compared with static threshold strategies. However, frequent VM migrations may introduce additional overhead and temporary performance degradation.

Huang et al. (2021) developed a deep neural network-based resource prediction model for proactive VM allocation in cloud environments. The proposed system uses historical workload data to train neural networks capable of predicting CPU and memory demands for upcoming workloads. Based on these predictions, the system proactively allocates virtual machines to meet anticipated resource requirements. The experimental evaluation showed that the model improved prediction accuracy and reduced resource wastage in cloud data centres. Nevertheless, the authors noted that deep neural network models require extensive training datasets to achieve optimal performance.

Sinha et al. (2022) introduced a hybrid metaheuristic algorithm for energy-efficient VM placement that combines Ant Colony Optimization (ACO) with Genetic Algorithms. The proposed algorithm evaluates multiple resource parameters such as CPU load, memory usage, and network bandwidth when determining VM placement decisions. The hybrid approach allows the system to explore a larger solution space and identify optimal allocation strategies. Simulation results indicated improved load balancing and reduced energy consumption compared with conventional VM allocation techniques. However, the hybrid optimization algorithm required longer computation times due to its complex search process.

Dasgupta et al. (2023) proposed a capsule neural network-based resource prediction framework for cloud data centres. Capsule networks were used to capture hierarchical relationships between resource utilization features such as CPU usage, memory demand, and storage consumption. The model demonstrated higher prediction accuracy compared with traditional convolutional neural networks. By integrating capsule networks with proactive auto-scaling mechanisms, the framework improved resource allocation efficiency and reduced energy consumption in cloud infrastructures. However, capsule networks require higher computational resources during the training phase.

Ahmed et al. (2023) developed an attention-based deep learning framework for energy-efficient VM allocation in cloud data centres. The system incorporates an attention mechanism that identifies the most influential workload features affecting resource consumption. By focusing on these critical features, the model improves prediction accuracy and enables more efficient VM allocation strategies. Experimental results showed significant improvements in resource utilization and energy efficiency compared with conventional machine learning models. Nevertheless, the framework requires high computational power and large datasets to train the attention-based neural networks effectively.

Comparative Table

Study	Author	Technique / Model	Objective	Key Contribution	Limitation
1	Saxena & Singh	Multi-resource neural network	Proactive auto-scaling	Improves resource prediction accuracy	Training complexity
2	Alharbi et al.	Energy-efficient VM placement	Energy reduction	Reduces power consumption in fog-cloud systems	Complex architecture
3	Katal et al.	Energy optimization techniques	Cloud energy management	Comprehensive energy efficiency strategies	Survey-based
4	Garg et al.	Particle swarm optimization	VM allocation optimization	Reduces SLA violations	Parameter tuning required
5	Arroba et al.	Metaheuristic thermal optimization	Cooling and computing efficiency	Improves energy efficiency	Complex modelling

6	Beloglazov et al.	Dynamic VM consolidation	Energy-efficient cloud management	Reduces active servers	Migration overhead
7	Mishra & Sahoo	ML-based auto-scaling	Workload prediction	Improves scaling accuracy	Requires retraining
8	Zhao et al.	Deep reinforcement learning	VM placement	Adaptive learning-based allocation	Training cost
9	Zhang et al.	LSTM-based prediction	Multi-resource forecasting	Improves SLA compliance	High computation
10	Wang et al.	Attention neural network	Resource prediction	Improves workload forecasting	Data intensive
11	Xu et al.	Energy-aware scheduling	Energy reduction	Improves workload classification	Requires accurate profiling
12	Kumar & Kumar	GA-PSO hybrid optimization	VM placement	Improves resource allocation	Algorithm complexity
13	Chen et al.	CNN-based prediction	Auto-scaling	Improves prediction accuracy	Training overhead
14	Patel et al.	VM consolidation algorithm	Energy efficiency	Reduces idle server energy	Migration cost
15	Ahmed et al.	Reinforcement learning scheduling	Multi-resource allocation	Adaptive scheduling	Training time
16	Li et al.	Heuristic VM migration	Load balancing	Improves utilization	Network overhead
17	Roy et al.	ML workload prediction	Auto-scaling	Proactive resource allocation	Data dependency
18	Singh et al.	Load balancing algorithm	Cloud performance	Improves throughput	Scalability issues
19	Khan et al.	Swarm intelligence optimization	VM allocation	Reduces operational cost	Parameter sensitivity
20	Luo et al.	LSTM + Attention model	Resource forecasting	Improves prediction accuracy	Computational complexity
21	Nguyen et al.	Workload consolidation	Energy management	Reduces idle servers	Monitoring dependency
22	Das et al.	Regression-based auto-scaling	Cloud scaling	Improves response time	Prediction accuracy issues
23	Gupta et al.	Task scheduling algorithm	Energy efficiency	Improves resource utilization	Processing overhead
24	Bansal et al.	Multi-objective optimization	VM placement	Balances energy and reliability	High computation
25	Reddy et al.	RNN workload prediction	Auto-scaling	Reduces SLA violations	Large datasets required
26	Sharma et al.	Dynamic threshold VM allocation	Energy optimization	Improves utilization	Migration overhead
27	Huang et al.	Deep neural network prediction	Resource forecasting	Reduces resource wastage	Large training datasets
28	Sinha et al.	ACO-GA hybrid algorithm	VM placement	Improves load balancing	Long computation
29	Dasgupta et al.	Capsule neural networks	Resource prediction	Improves hierarchical feature learning	High training cost
30	Ahmed et al.	Attention-based deep learning	VM allocation	Improves prediction accuracy	High computational requirements

Conclusion

Cloud computing has become a fundamental infrastructure for delivering scalable computing services, hosting large-scale applications, and managing massive volumes of data across distributed environments. As cloud adoption continues to grow, cloud data centres must handle increasingly dynamic workloads while maintaining high performance, reliability, and energy efficiency. One of the most critical challenges faced by cloud service providers is efficient resource management, particularly in the areas of virtual machine allocation and auto-scaling. Improper resource allocation can lead to underutilized servers, excessive energy consumption, and violation of service level agreements (SLAs). Therefore, developing intelligent frameworks for proactive auto-scaling and energy-efficient VM allocation has become a major research focus in recent years.

This review paper analysed recent research studies published between 2020 and 2023 that focus on proactive auto-scaling and energy-efficient VM allocation techniques for cloud data centres. The literature review examined thirty research studies covering various approaches including heuristic algorithms, metaheuristic optimization methods, machine learning techniques, and advanced deep learning models. The findings indicate that traditional heuristic-based VM allocation methods are gradually being replaced by intelligent resource management techniques that can dynamically adapt to changing workload conditions.

Machine learning and deep learning models have shown significant potential in predicting resource utilization patterns and enabling proactive auto-scaling strategies. Predictive models such as recurrent neural networks, convolutional neural networks, and long short-term memory networks can analyse historical workload data to forecast future resource demands. These predictions allow cloud management systems to allocate virtual machines in advance, thereby preventing performance degradation and reducing SLA violations. In addition, attention mechanisms and capsule neural networks have demonstrated improved capability in capturing complex relationships between multiple resource parameters, such as CPU usage, memory consumption, and network bandwidth.

References

Alharbi, F., et al. (2020). Energy-efficient virtual machine placement in cloud-fog environments. *Future Generation Computer Systems*, 107, 102–112. https://doi.org/10.1016/j.future.2020.01.032

Arroba, P., et al. (2023). Metaheuristic optimization for energy management in cloud data centres. *IEEE Access*, 11, 35001–35015. https://doi.org/10.1109/ACCESS.2023.3262204

Beloglazov, A., Abawajy, J., & Buyya, R. (2020). Energy-aware resource allocation heuristics for efficient management of data centers. *Future Generation Computer Systems*, 28(5), 755–768. https://doi.org/10.1016/j.future.2011.04.017

Bansal, S., Kumar, R., & Singh, D. (2022). Multi-objective VM placement in cloud computing using evolutionary algorithms. *Applied Soft Computing*, 115, 108153. https://doi.org/10.1016/j.asoc.2021.108153

Chen, X., Zhang, Y., & Wang, J. (2021). Deep learning-based predictive auto-scaling framework for cloud applications. *Journal of Cloud Computing*, 10(1), 1–14. https://doi.org/10.1186/s13677-021-00235-8

Das, S., et al. (2021). Machine learning-based proactive auto-scaling for cloud environments. *IEEE Access*, 9, 150123–150134. https://doi.org/10.1109/ACCESS.2021.3124576

Dasgupta, S., et al. (2023). Capsule neural network-based resource prediction for cloud computing. *Computers & Electrical Engineering*, 105, 108493. https://doi.org/10.1016/j.compeleceng.2022.108493

Garg, S., et al. (2023). Particle swarm optimization-based VM allocation for cloud data centres. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5). https://doi.org/10.4108/eai.18-11-2022.2327125

Huang, Q., et al. (2021). Deep neural network-based resource prediction for cloud auto-scaling. *Future Generation Computer Systems*, 118, 199–210. https://doi.org/10.1016/j.future.2021.01.028

Khan, S., et al. (2022). Swarm intelligence-based VM placement for energy-efficient cloud computing. *Applied Soft Computing*, 118, 108512. https://doi.org/10.1016/j.asoc.2022.108512

Kumar, P., & Kumar, R. (2021). Hybrid GA–PSO algorithm for energy-efficient VM allocation in cloud computing. *Journal of Supercomputing*, 77(8), 8345–8363. https://doi.org/10.1007/s11227-020-03612-9

Li, Y., et al. (2020). Heuristic-based VM migration strategies for energy-efficient cloud computing. *IEEE Transactions on Cloud Computing*, 8(2), 558–570. https://doi.org/10.1109/TCC.2017.2769673

Luo, Z., et al. (2023). Attention-based neural networks for cloud resource prediction. *IEEE Transactions on Network and Service Management*, 20(1), 45–57. https://doi.org/10.1109/TNSM.2022.3207814

Mishra, A., & Sahoo, B. (2021). Intelligent workload prediction for proactive cloud resource scaling. *Journal of Systems and Software*, 179, 111014. https://doi.org/10.1016/j.jss.2021.111014

Nguyen, T., et al. (2020). Workload consolidation techniques for energy-efficient cloud data centres. *Future Generation Computer Systems*, 102, 331–345. https://doi.org/10.1016/j.future.2019.08.022

Patel, H., et al. (2022). Energy-efficient VM consolidation algorithms for cloud infrastructures. *IEEE Access*, 10, 41235–41247. https://doi.org/10.1109/ACCESS.2022.3162345

Reddy, K., et al. (2023). Recurrent neural network-based workload prediction for cloud auto-scaling. *Computers & Electrical Engineering*, 106, 108512. https://doi.org/10.1016/j.compeleceng.2023.108512

Roy, A., et al. (2021). Predictive workload management for cloud computing using machine learning. *Journal of Cloud Computing*, 10(1), 1–17. https://doi.org/10.1186/s13677-021-00240-x

Saxena, D., & Singh, A. (2021). Proactive auto-scaling framework using multi-resource prediction for cloud computing. *Neurocomputing*, 427, 97–109. https://doi.org/10.1016/j.neucom.2020.11.066

Sharma, S., et al. (2020). Dynamic threshold-based VM allocation for energy-efficient cloud computing. *Future Generation Computer Systems*, 102, 264–276. https://doi.org/10.1016/j.future.2019.08.015

Singh, P., et al. (2021). Load balancing strategies for VM allocation in cloud data centres. *IEEE Access*, 9, 101234–101245. https://doi.org/10.1109/ACCESS.2021.3098456

Sinha, R., et al. (2022). Hybrid ACO-GA algorithm for energy-efficient VM placement. *Applied Soft Computing*, 120, 108690. https://doi.org/10.1016/j.asoc.2022.108690

Wang, Y., et al. (2023). Attention-based resource prediction for cloud computing environments. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245123

Xu, J., et al. (2020). Energy-aware scheduling algorithms for cloud computing systems. *Future Generation Computer Systems*, 108, 105–115. https://doi.org/10.1016/j.future.2020.02.033

Zhao, L., et al. (2021). Reinforcement learning-based VM placement for cloud resource optimization. *IEEE Access*, 9, 15872–15885. https://doi.org/10.1109/ACCESS.2021.3051542

Zhang, Y., et al. (2022). LSTM-based multi-resource prediction for proactive cloud scaling. *Future Generation Computer Systems*, 129, 220–231. https://doi.org/10.1016/j.future.2021.11.019

Gupta, R., et al. (2021). Energy-efficient task scheduling in cloud computing environments. *Journal of Network and Computer Applications*, 190, 103132. https://doi.org/10.1016/j.jnca.2021.103132

Ahmed, M., et al. (2022). Reinforcement learning-based scheduling for cloud resource management. *Computers & Security*, 117, 102698. https://doi.org/10.1016/j.cose.2022.102698

Chen, L., et al. (2021). Deep neural networks for resource allocation in cloud infrastructures. *Information Sciences*, 524, 234–248. https://doi.org/10.1016/j.ins.2020.03.031

Alharbi, F., et al. (2020). Energy-efficient resource allocation strategies in cloud computing. *Journal of Cloud Computing*, 9(1), 1–12. https://doi.org/10.1186/s13677-020-00185-2