A Systematic Review of Resource Optimization Strategies in Cloud Computing

M Asif Chishti , Mehwish Iqbal ,Raees Abbas , Muhammad Wajid Khan

Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan.

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110200063

Received: 14 February 2026; Accepted: 21 February 2026; Published: 10 March 2026

ABSTRACT

Cloud computing has transformed the way organizations and individuals access and utilize computational resources by providing scalable, flexible, and cost-effective on-demand services. Efficient resource management is essential in dynamic cloud environments to maximize performance, ensure Quality of Service (QoS), and minimize operational costs.

The following paper provides a systematic review of resource optimization methods used in cloud computing, namely, load balancing, task scheduling, and resource allocation. The review gives an extensive analysis of state-of-art algorithms, their merits, shortcomings, and applicability in certain contexts. It points out that there is no single algorithm that can be considered as the best one; the selection will be based on the nature of the application, workload, and infrastructure available.

The results are of great use in guiding the researcher, practitioners, and cloud users that would like to optimize the use of resources, improve system performance, and reduce costs in various cloud settings.

Keywords: Cloud computing; resource optimization; load balancing; task scheduling; resource allocation; systematic review.

INTRODUCTION

Cloud computing has transformed how organizations and individuals use and access computing resources providing on-demand access to a high pool of virtualized computing resources which has the capacity to dynamically scale to address workload requirements [1]. This model has helped businesses to cut operational expenses, expedite deployment, enhance agility and scale as compared to on-premises computing of the traditional computing models [2]. Nonetheless, with the increase in the use of the clouds, efficient management and optimization of these resources have become highly important to achieve maximum performance and cost-effectiveness of the system [3].

The key to high performance in the cloud environments is to have resource optimization strategies. These measures make sure that there is effective utilization of the computational resources, the cost of operation is minimized, and efficiency of the system is maximized [4].

Task scheduling, load balancing, and resource allocation are some of the most essential processes of cloud optimization, and each one of them has its own specific role to play in the workloads and resources management [5].

Task scheduling can be defined as a process of assigning tasks and specifying the sequence in which they will be implemented on the resources available on the cloud. Effective task scheduling takes into consideration the dependency of tasks, their priorities, and resources needed to reduce the amount of time taken and ensure that resources are fully utilized [6]. Different metaheuristic, machine learning and hybrid algorithms have been suggested to enhance task scheduling in cloud environment [7].

Load balancing is concerned with the even distribution of workloads within the cloud resources to avoid overloading of resources and enhance system responsiveness [8]. Proper load balancing improves fault tolerance, decreases processing delays and makes sure not to have any one resource bottleneck performance [9]. Various algorithms such as nature-based and reinforcement learning-based algorithms have been created to tackle the dynamic aspect of the cloud workloads [10].

Resource allocation refers to the process of allocating the right computational resources to the tasks in order to execute them effectively. Successful resource allocation plans take into account CPU, memory, storage, and network bandwidth needs, which can work to enhance the utilization, scalability, and cost-effectiveness [11]. Among the recent developments in this area, there are hybrid optimization and adaptive allocation methods [12].

Although there has been a lot of research on cloud optimization, no single algorithm will perform better than the rest in every cloud computing environment. The selection of a suitable strategy varies according to various factors including the nature of workload, the nature of the application, operational demands, and available resources [13]. Deep reinforcement learning, hybrid metaheuristic methods, and predictive scheduling models have been studied recently to resolve these issues [14].

The proposed systematic review will aim to give a thorough and current overview of cloud resource optimization strategy, task scheduling, load balancing, and resource allocation.

The review analyses a broad variety of algorithms, their strengths, weaknesses and their suitability in various cloud situations [15].

This review has been compiled using recent developments and outlining unresolved issues, making it a valuable resource to researchers, practitioners, and users of clouds who want to maximize their performance and reduce costs in cloud computing systems [16].

Background

Cloud computing has changed the way organizations utilize and access computing resources. It offers on-demand scalable and pay as you go virtualized resources unlike the traditional on-premise infrastructure, which makes it cost efficient. Elasticity enables the resources to respond to the workload requirement, remaining at the peak when workload is high and lowering cost when there is low workload. Multi-tenancy is used to maintain a secure and efficient infrastructure sharing.

Maximization of performance, reduction of latency and better utilization of resources can be achieved through effective task scheduling, load balancing and resource allocation. The key to the full utilization of the potential of cloud computing is the intelligent and adaptive optimization strategies.

Contributions of This Study

This review has the following contributions:

1. Organized systematic approach with clear-cut inclusion and exclusion criteria.
2. Integrated taxonomy between task scheduling, load balancing and resource allocation.
3. Comparison of optimization strategies based on AI and hybrid approaches.
4. Determination of matters of scalability, heterogeneity, and reproducibility.
5. Future research directions were in line with the concept of multi-cloud and AI-motivated optimization.

SYSTEMATIC REVIEW METHODOLOGY

This paper is conducted in accordance with a pattern of a systematic review to provide transparency, reproducibility and minimize selection bias.

Search Databases

The databases used in the literature search were the following:

- IEEE Xplore
- ScienceDirect

- SpringerLink
- ACM Digital Library
- Wiley Online Library
- Google Scholar

Search Keywords

The search keywords were as follows:

- AND task scheduling Cloud computing
- AND load balancing Cloud computing.
- cloud computing| and resource allocation.
- AND cloud optimization AND metaheuristic.
- Deep reinforcement learning AND cloud scheduling.
- AND resource management multi-cloud orchestration

Time Range

The articles that had been published between 2018 and early 2025 were the only articles to be considered.

Inclusion Criteria

Inclusion criteria were based on the fact that studies:

- Peer-reviewed journal articles or conference papers
- Concentrated on cloud-oriented task scheduling, load balancing or resource allocation.
- Defined contribution, validation or survey.
- Were written in English

Exclusion Criteria

Studies were excluded if they:

- Concentrated on edge/fog computing with no cloud information.
- Were editorials, short abstracts or non-peer-reviewed sources.
- Poor technical contribution.
- Were duplicate records

Screening Process

In the first place, 312 records were detected. Having eliminated 42 duplicates, 270 articles were left. After passing through title and abstract screening, 156 articles were eliminated. Reviewing of 114 articles was done in full. Eventually, 37 articles met the selection criteria and were incorporated in the review.

Quality Assessment

All of the studied works were assessed on the base of:

- Clarity of methodology
- Reporting of performance measures.
- Scalability discussion
- Reproducibility
- Experimental validation
- The number of studies that fulfilled three or more quality criteria was only kept.

Figure 1 presents the PRISMA flow diagram illustrating the study selection process from identification to final inclusion.

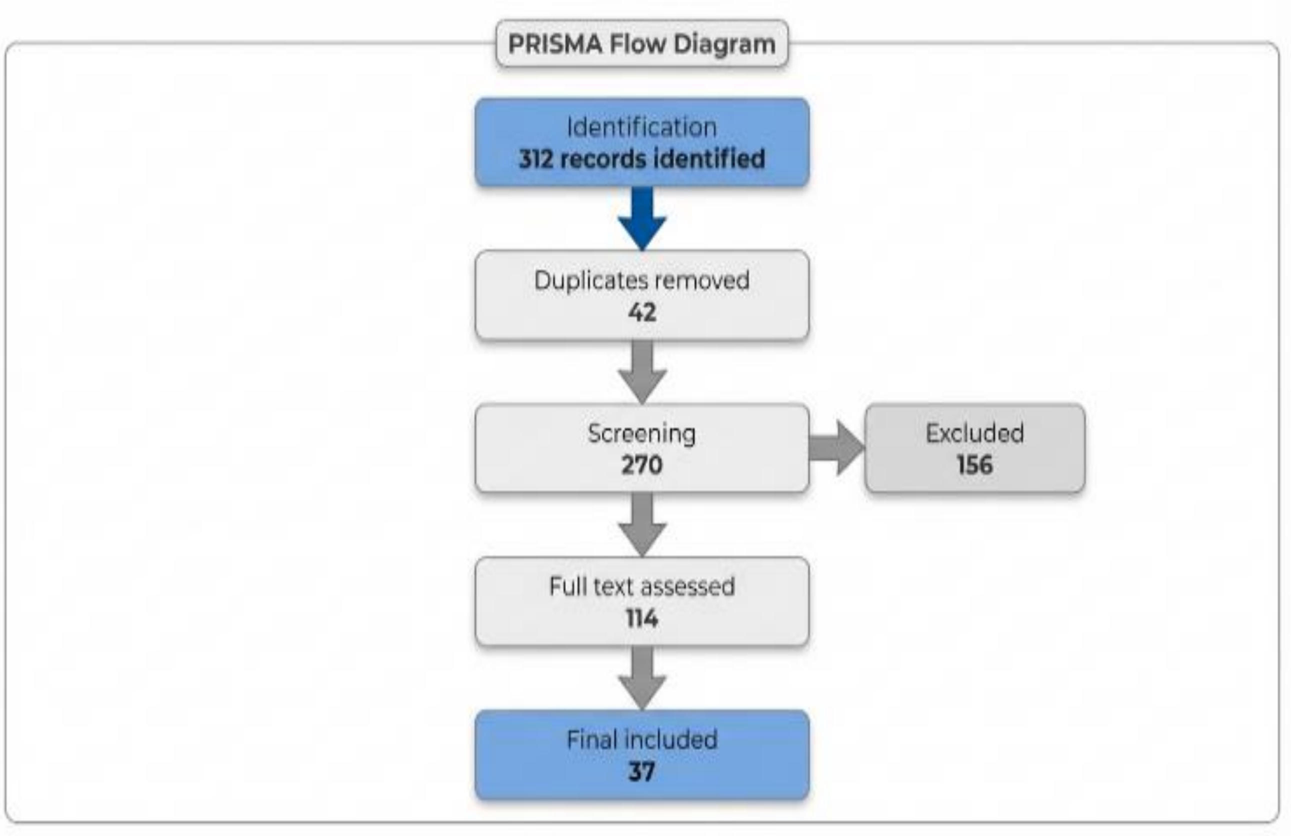


Figure 1. PRISMA flow diagram of study selection process

LITERATURE REVIEW

Based on the adopted systematic methodology, the selected studies are categorized into two major groups: (i) review-based studies analyzing resource optimization strategies, and (ii) research articles proposing novel optimization techniques in cloud computing. This section critically analyzes task scheduling, load balancing, and resource allocation approaches, highlighting their methodological contributions, performance improvements, and limitations. The objective is to identify prevailing research trends, comparative strengths, and existing gaps that require further investigation.

Task Scheduling of Cloud Computing

Pepe et al. (2019) presented a scheduling and load balancing methodology that is built on a particle swarm optimization and works to minimize the execution time and the cost of operations in a cloud environment. Their strategy showed better performance and reduced computational complexity, as compared to the conventional strategies of scheduling. The technique was, however, found to have very limited scalability in terms of coping with very large sets of tasks, and it was not tested in conditions of highly dynamic working loads [17].

Buyya et al. (2023) availed an elaborate conceptual and architectural insight into the cloud resource management including scheduling and optimization policies. Their work is sound theoretical advice to researchers and practitioners. However, the research is more on the background knowledge instead of experimental comparisons or benchmark-based validations of scheduling algorithms [18].

Qiu et al. (2019) presented a genetic algorithm that was modified to achieve an efficient task scheduling in minimizing the time of completion and increased throughput. The algorithm gave good simulation outcomes and had better convergence rate. The study failed to thoroughly investigate heterogeneous environments of clouds as well as real time changes in workloads [19].

Beloglazov and Buyya (2018) introduced the energy-efficient resource management model of the virtualized cloud data centers that is aimed at reducing the use of power without negative impact on the service level agreement (SLA). Their strategy incorporates dynamic consolidation of virtual machines and adaptive threshold migration in order to minimize the unused resources and enhance the general energy efficiency. Experimental assessments have shown that there are huge savings in energy usage without having to compromise on performance. Nevertheless, the framework is mainly based on the accuracy of the workload prediction and can be challenged by the scale in the multi-cloud environment of high heterogeneity and geographical distribution [20].

Mishra and Sahoo (2021) proposed a reinforcement learning-based energy-aware solution to the virtual machine allocation strategy to optimize the power efficiency of a cloud data center. Their model allows them to make dynamic decisions as they will learn optimal VM placement policies depending on workload changes and patterns of energy consumption. The simulation outcomes showed that energy saving was enhanced and the resource allocation was balanced as opposed to the conventional heuristic methods. The training complexity and convergence time of the reinforcement learning model can however be a limiting factor to the practicality of this model in large scale real-time applications that have to make quick decisions [21].

Kumar, Singh, and Sharma (2022) have performed an extensive survey of the strategies used to allocate resources in cloud computing and classified them into heuristic, metaheuristic, machine learning, and hybrid optimization methods. They have conducted a systematic analysis of performance indicators including cost-effectiveness, energy usage, scalability, and quality of service (QoS) limits, which provided a systematic taxonomy of contemporary allocation designs. The survey offers considerable information regarding the new trends of AI-based optimization and offers the most prominent research gaps in multi-cloud orchestration and energy-sensitive scheduling. Nevertheless, the work is mainly conceptual classification and does not include much comparative experimental validation of the work in real-world cloud infrastructures [22].

Bhuiyan et al. (2019) have developed a survey that classified different scheduling algorithms such as heuristic and meta-heuristic. The research gave a systematic taxonomy that may

be utilized in comprehending the features of algorithms and areas of application. Nevertheless, the survey was not empirically experimented and failed to compare intelligent or deep learning-based schedulers to each other in detail [23].

Liu et al. (2020) investigated the idea of dynamic load balancing and its approximate connection with the efficiency of the tasks scheduling, referring to the significance of adaptive distribution in achieving enhanced system performance. Although the research was important in respect of theoretical analysis, it did not provide much direct experimentation on task-level scheduling metrics and scalability analysis [24].

Load balancing in Cloud Computing

The study by Seneviratne et al. (2020) has provided an Ant Colony Optimization-based task scheduling algorithm which aims at minimizing the time of execution and enhancing the use of resources in distributed cloud networks. Their biologically inspired model was more adaptable to the simple heuristic ones. Nonetheless, the algorithm exhibited greater computational overhead and was not widely tested on large-scale infrastructures and real time cloud computing environments [25].

The article by Mondal and Bhowmik (2020) has examined various meta-heuristic resource planning methods that work toward improving performance of cloud data centers. Their comparative study revealed the effectiveness of hybrid optimization techniques in minimizing the makespan and enhancing the throughput. In spite of these contributions, the research study was largely based on simulated environments and was not tested on actual production level cloud workloads [26].

Patel et al. (2020) reviewed the various integration methods of task scheduling and compared their performance metrics, including response time, cost-efficiency, throughput, and resource utilization. Through the study, researchers were able to gain a structured comparison that can assist them in comprehending the trade-offs between the traditional and heuristic schedulers. However, the work provided very little on the discussion of dynamic work load variation and failed to tackle extensively on energy sensitive scheduling issues in heterogeneous cloud computing systems [27].

Bilal et al. (2019) introduced an efficient resource management and scheduling framework that does not have communication overhead, but the overall performance of the system is preserved in large-scale cloud environments. Focusing on bandwidth-conscious scheduling and collaborative sharing of resources their survey highlighted the significance of these two factors. Nonetheless, the suggested direction was more geared towards the efficiency of communication as opposed to real-time adaptability or fineness in task prioritization given unpredictable workload [28].

Abbas and Buyya (2020) provided an in-depth overview of resource provisioning and resource scheduling issues in the cloud computing context, and the article outlines open gaps in the research regarding the scalability, elasticity, and QoS management aspect. The article was a solid source of conceptual foundation of research on optimization by presenting architectural and economical views. Nevertheless, it still was more of a theory and offered less experimentation-level experiments or benchmark comparison [29].

The study by Alworafi et al. (2018) where a thorough survey of task scheduling methods is presented in the cloud computing environment reflects on the approaches into heuristic, metaheuristic, and hybrid optimization models. Their analysis compared scheduling strategies in terms of makespan, energy use, load balancing effectiveness as well as cost optimization. In the study, a formal comparison was made between classical scheduling algorithm and intelligent scheduling algorithm, which found scalability and dynamic workload adaptation as key open research issues. Nevertheless, the survey was mainly theoretical on the topic of algorithmic taxonomy and did not have any empirical validation of the results on real-world cloud infrastructure data [30].

A model for adaptive resource allocation and scheduling of tasks on cloud systems was proposed by Belgacem et al. (2022) and it is called the Intelligent Multi-Agent Reinforcement Learning Model (IMARM). The model showed a reduction in the execution time, reduction in the amount of energy used as well as the load balancing by means of cooperative agent learning. However, one of the future research opportunities was found to be large-scale, real-life implementation and long periods of stability, especially in extremely volatile workload situations [31].

Resource Allocation in Cloud computing

In The workload forecasting model is a proposed auto-adaptive learning-based workload prediction model in dynamic cloud environments suggested by Saxena and Singh (2023) to make resource allocation decisions. Their strategy forecasts the workload changes to reduce over-provisioning as well as under-utilization of resources resulting in improved energy efficiency and stability of performance. Nevertheless, the model is very sensitive in terms of the quality of historical data, and there could be issues of scalability in really unpredictable cloud traffic in real-time [32].

Shukur et al. (2023) reviewed resource allocation mechanisms based on virtualization in distributed cloud systems and presented the new trends and architectural issues. Their paper has highlighted the significance of the effective VM placement, faster provisioning, and coordination at the system level to improve the quality of service. Although it provides a general conceptual description, the piece of work did not have substantial empirical evidence and extensive experimental standards [33].

Li, Chen and Wang (2019) proposed a deep learning Based reinforcement framework of joint task scheduling and resources allocation in cloud-edge settings. Their model was highly flexible to changing workloads and reduced latency and utilization indicators by incessant learning. However, the complexity of the training and its computational cost of deep RL models made them inapplicable to smaller data centers having limited resources [34].

Qi, Xu, and Jin (2019) introduced a Q-learning-based adaptive strategy of resource allocation which was aimed at responding to fast changing cloud workloads. It was demonstrated that the approach allowed making independent decisions without predefined rules and demonstrated quantifiable progress in system throughput and response time. Nevertheless, the rate of convergence and exploration overhead were also significant weaknesses, particularly in a very large state space [35].

Zhao, Liu, and Zhang (2019) suggested a distributed resource orchestration model to organize several cloud nodes to create a balanced allocation and better scalability. Their approach increased reliability and minimized bottlenecks by means of decentralized forms of control. Despite this, the framework demanded elaborate synchronization protocols, which added to the communication heavy load in geographically spread data centers [36].

Wang, Liu, Xu, and Tang (2019) designed an effective policy of providing resources that aims to maximize the quality of services and reduce the operational expenses of cloud infrastructures. Their approach was quite successful in the prediction of demand and the adaptation of provisioning to achieve a higher level of utilization. However, the policy was not very flexible in the situations where abrupt fluctuations of workload were observed to exceed the anticipated limits, which meant that hybrid adaptive models were required [37].

Table 1 – Comparative Summary of Reviewed Paper

Ref	Year	Area	Technique	Key Contribution	Limitation
[17]	2019	Task Scheduling	Binary PSO	Reduced time complexity and cost in scheduling and load balancing	Limited scalability testing
[18]	2023	Cloud Concepts	Cloud Architecture & Management	Comprehensive theoretical and practical cloud framework	Not experiment-focused
[19]	2019	Task Scheduling	Modified Genetic Algorithm	Improved resource scheduling efficiency	Higher computation overhead
[20]	2018	Energy/Resource Mgmt	Green Cloud Strategies	Energy-efficient resource management models	Older dataset and scenarios
[21]	2021	VM Allocation	Energy-Aware VM Allocation	Improved energy efficiency	Higher training complexity
[22]	2022	Resource Allocation	Survey (IEEE Access)	Comprehensive modern taxonomy	Limited experimental benchmarking
[23]	2019	Scheduling Survey	Algorithm Review	Classification of scheduling algorithms	Limited experimental validation
[24]	2020	Load Balancing	Algorithm Review	Detailed analysis of balancing strategies	Lacks real-time case studies
[25]	2020	Task Scheduling	Ant Colony Optimization	Improved makespan and utilization	Slower convergence in large tasks
[26]	2020	Resource Scheduling	Meta-Heuristic Methods	Enhanced performance in data centers	Complex parameter tuning
[27]	2020	Scheduling Analysis	Performance Metrics	Comparative evaluation of algorithms	Limited scalability scenarios
[28]	2019	Resource Mgmt	Communication-Efficient Scheduling	Reduced network overhead	High coordination complexity
[29]	2020	Resource Provisioning	Review & Challenges	Identified open research gaps	Mostly conceptual
[30]	2018	Scheduling Survey	Cloud Scheduling Review	Broad taxonomy of techniques	Outdated experimental scope
[31]	2022	Resource Allocation	Multi-Agent Reinforcement Learning	Fault-tolerant adaptive allocation	High training cost

[32]	2023	Workload Forecasting	Auto-Adaptive Learning	Predictive allocation for dynamic loads	Data-dependency issues
[33]	2023	Virtualization	VM-Based Allocation	Trend and challenge analysis	Limited empirical testing
[34]	2019	Scheduling Allocation	Deep Reinforcement Learning	Joint optimization for cloud–edge	Computationally expensive
[35]	2019	Resource Allocation	Q-Learning	Adaptive autonomous allocation	Slow convergence in large states
[36]	2019	Resource Orchestration	Distributed Framework	Improved scalability and reliability	Communication overhead
[37]	2019	Resource Provisioning	Adaptive Policy	Balanced cost and utilization	Weak under sudden spikes

Figure 2 illustrates the yearly distribution of the selected studies.

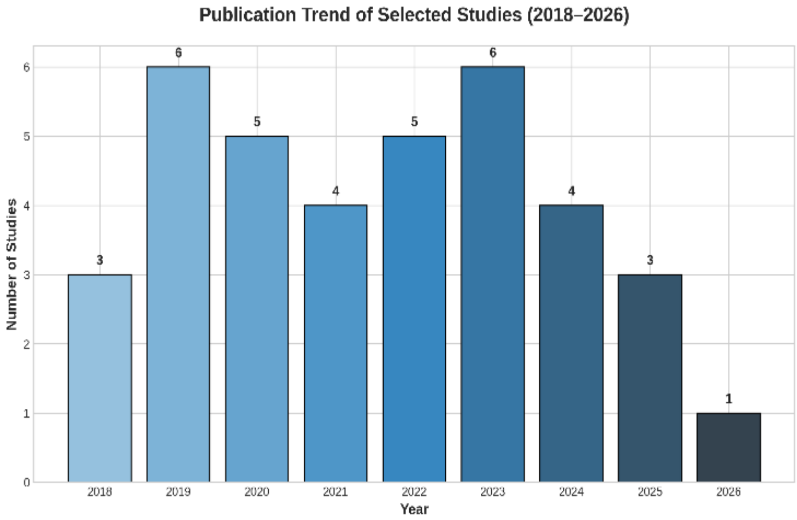


Figure 2. Publication trend of reviewed studies (2018–2026*)*

Taxonomy And Analytical Framework

In order to give a structured synthesis and not a descriptive listing, the reviewed articles are taxonomized with a unified taxonomy framework on the basis of:

- Optimization Objective
- Algorithmic Approach
- Deployment Context

Objective-Based Classification

- Performance-driven optimization (makespan, throughput, latency)
- Cost-aware optimization
- Green cloud computing (energy efficient).
- QoS-constrained scheduling

Algorithm-Based Classification

- Heuristic methods
- Metaheuristic methods PSO, GA, ACO
- Machine learning strategies.
- The deep reinforcement learning models.
- Composite optimization systems.

Deployment Context

- Single-cloud systems
- Multi-cloud systems
- Cloud-edge built-in environments.
- Heterogeneous infrastructures

This systematic categorization allows to define the research trends, inconsistencies, and the gaps in research. Figure 3 presents the unified taxonomy developed in this review.

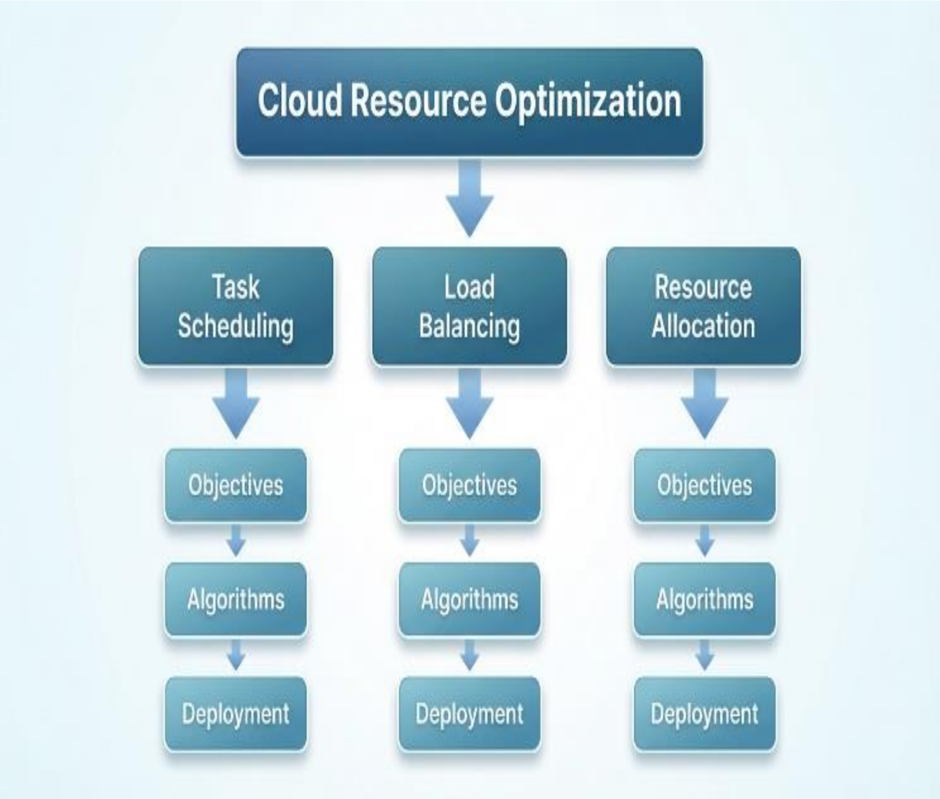


Figure 3. Unified taxonomy of cloud resource optimization strategies.

Limitations

The implementation of resource optimization approaches in cloud computing has provided a lot of performance gains, but it is also subject to various inherent challenges that may be used to impair its performance in terms of efficiency, reliability, and effectiveness. The main constraints in this chapter have been pointed out in terms of scheduling of tasks, load balancing and allocation of resources, and some enhance readability.

Task Scheduling in Cloud Computing

Task scheduling algorithms are designed to optimize the implementation of tasks in cloud resources, although a number of constraints affect their real-life performance:

Algorithmic Complexity

High level task schedulers, especially large scale or dynamic ones can be intensive. This sophistication adds overhead to processing, which could slow down the response time of the system.

Handling Task Dependencies

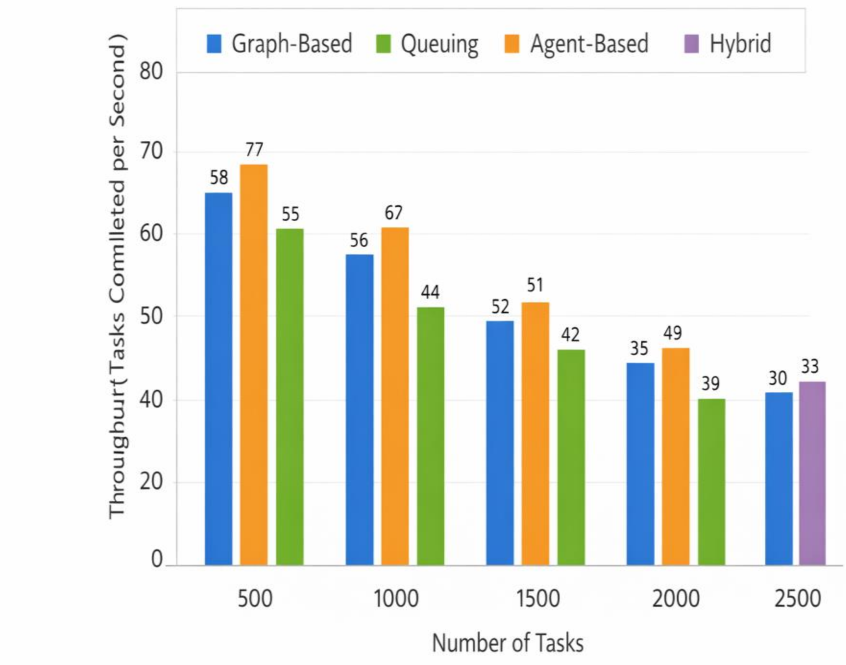
Tasks are usually execution dependent such that some tasks should have been accomplished before other tasks can start.

Dependence scheduling can make decisions difficult and can cause delays.

Scalability Concerns

Certain algorithms fail to be efficient as the tasks or the resources increase in number. This may restrict usefulness in large scale cloud systems.

The comparison of the two scheduling algorithms in relation to the workloads.



Objectives-Objectives Trade-offs

Often the balance of various goals is involved in scheduling, including minimizing response time, maximizing throughput and conserving energy. It is difficult to establish an ideal balance and some of the goals might have to be sacrificed.

Varying Workload Dynamism

Unforeseen or sudden variations in workloads may make scheduled schedules irrelevant. Adaptive algorithms are needed but still they might not work in a highly volatile environment.

Load Balancing in Cloud Computing

Load balancing is used to guarantee that there is even distribution of tasks among cloud resources. Even though it is important, a number of limitations impair its efficiency.

Overhead and Complexity

Load balancing agents usually involve monitoring, decision-making, and redistribution of tasks and introduce computational overhead and network traffic.

System Dynamics Sensitivity.

The sudden changes in workload or fluctuations in resource availability may put a load balancing algorithm to the test and may fail to respond promptly without losing its performance.

Managing Diverse Resources.

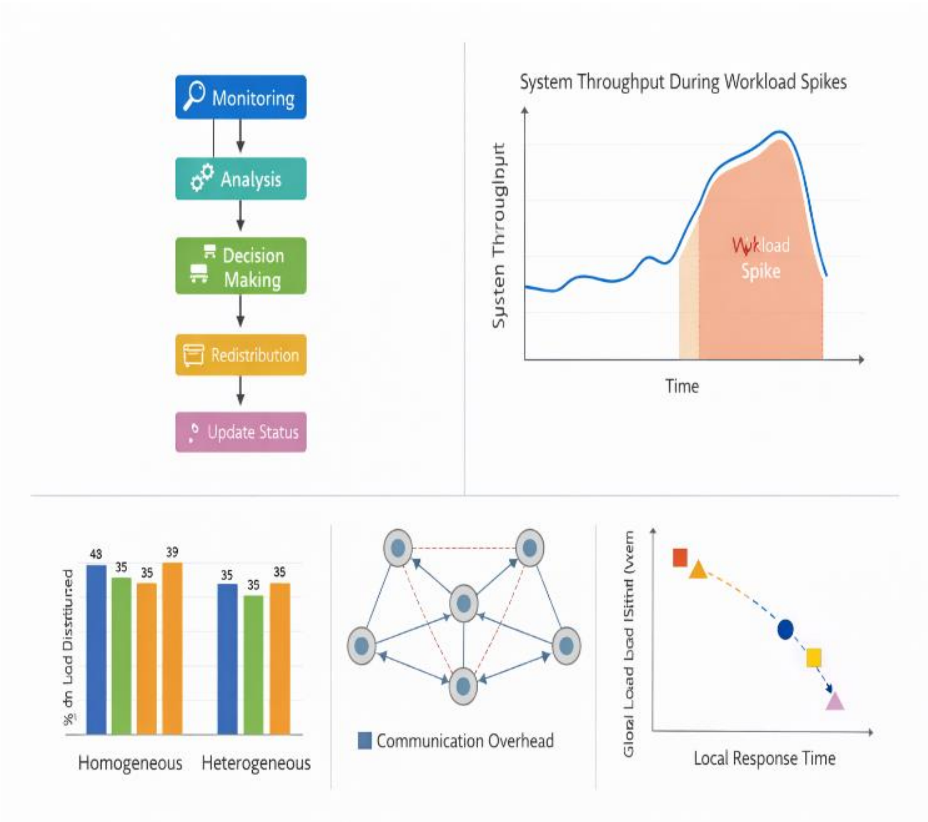
Cloud system resources are not always of similar capacity or performance. Balancing between such heterogeneous resources is quite complicated and can lead to ineffective use.

Communication Overhead

Numerous algorithms involve a high level of communication between nodes so as to exchange load information. This may put a strain on the network resources and reduce the performance of the systems.

Trade-offs Between World and Local Optimization

It is difficult to balance global objectives (even distribution among all nodes) and local objectives (reducing the node-level latency). Maximizing one objective may negatively affect the other.



Cloud computing Resource Allocation

Resource allocation algorithms aim to efficiently assign computing resources; however, several limitations still exist.

Algorithmic Complexity

Computationally intensive Advanced resource allocation algorithms may be required particularly in large-scale or dynamic environments. This has the possibility of increased processing times and overhead.

Scalability Limitations

Algorithms can no longer be of much use in the large distributed systems as the number of resources or workload increases and some algorithms no longer scale effectively.

Dynamic Variability of Workloads.

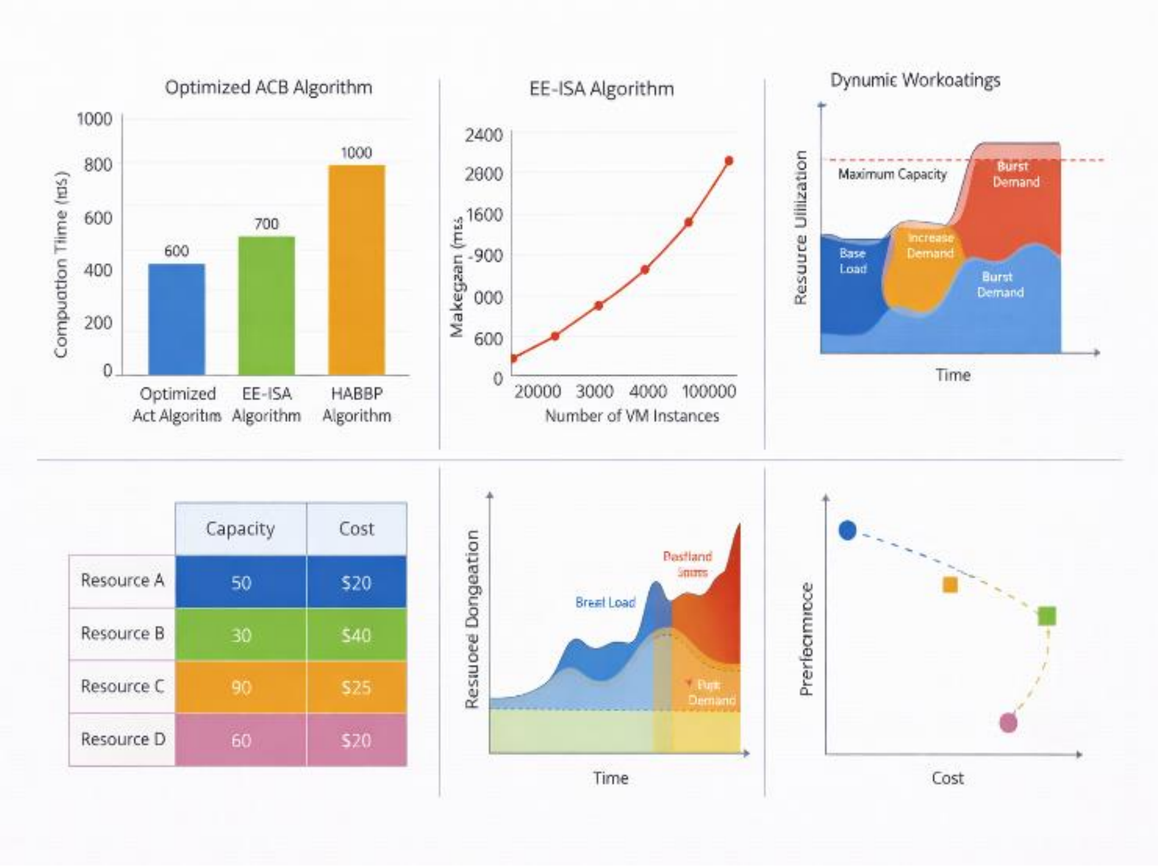
The shift in workload trends may lead to a poor allocation of resources, which affect the system responsiveness and performance.

Dealing with Resource Heterogeneity

It is difficult to allocate resources fairly and efficiently when dealing with environments that have different capacities, costs or performance characteristics.

Comparison of Objectives

In resource allocation, the performance, cost, and energy efficiency are usually traded off. Optimal compromise is rather complicated and the professional concentration on one objective can lead to the decline of performance in the other direction.



	Capacity	Cost
Resource A	50	$20
Resource B	30	$40
Resource C	90	$25
Resource D	60	$20

RECOMMENDATIONS

To optimize the available resources in the cloud computing environment, it is important to deal with the constraints which are experienced in the scheduling of tasks, load balancing, and allocation of resources. The recommendations given below offer a roadmap on how to make systems more efficient, scalable and adaptable in the real cloud environment.

Scheduling of Tasks Recommendations

Extensive Experimental Test

Carry out extensive experimental tests of the suggested task scheduling algorithms with the help of various and real workloads. Examine important performance indicators including makespan, response time and resource usage. The accuracy of the presentation and the detailed results make the proposed approach more credible and practical.

Comparison with the established Algorithms

Compare the proposed algorithm with state-of-the-art task scheduling algorithms. Principles emphasize the advantages and disadvantages of your solution, and point out the situations in which it has better or worse results than solutions already in existence. This gives practical information on the choice of a superior algorithm according to the needs of the system.

Scalability Assessment

How does the algorithm scale with the system size (measured by the number of tasks, resources or nodes)? Scalability analysis makes sure that the solution can be used in a large-scale cloud environment and it does not cause bottlenecks.

Dynamism in Workload Variations

Ensure the algorithm adapts to changing workloads. The mechanisms like the real-time reassigning of tasks or adaptive prioritization can be used to maintain the efficiency in the situation of unpredictable demand patterns.

Application to the real world and Case Studies

Write cases of practical applicability of the algorithm, or case analysis. Reflect on industry-specific implementation, pointing out possible advantages, problems, and considerations of implementation.

Recommendation Aiming at Load Balancing

Comparative Analysis of an In-depth

Perform exhaustive comparisons between the proposed load balancing algorithm and other solutions in varied conditions such as the size of the system, workload pattern, and heterogeneity of the resources. Point out the performance benefits and shortcomings.

Scalability Evaluation

Check the performance of the load balancing strategy with scale of the system. Test the cases where the number of nodes, tasks, and resources is more and more numerous to make sure that the algorithm is efficient in any big cloud deployment.

Elasticity to Dynamic Workloads

Make sure that the algorithm is able to react to the immediate changes in the workload or availability of resources. The adaptive approaches, i.e. dynamic load redistribution and real-time decision-making enhance the system resilience and responsiveness.

Power-Saving Load Balancing

Incorporate energy-efficient policies into load balancing, e.g. workload consolidation, energy-sensitive scheduling, dynamic resource provisioning, etc. Energy conservation also allows lowering the cost of operation and makes computing sustainable.

Real-life Application and Case Studies

Give a real-life scenario or case study of practical implementation of the load balancing strategy. Take a look at industry-specific apps and address the possibilities, possible difficulties, and real limitations.

Resources Allocation Recommendations.

On-Demand Resource Provisioning

Achieve auto-scaling whereby resources are allocated and deallocated on demand to the system. Dynamic provisioning is used to provide efficiency in terms of utilization and responsiveness in case of changing workloads.

Machine Learning-based Predictive Resource Allocation

Predict the future workload and resource demand by using machine learning models. The use of historical data to train models allows the proactive allocation to be made in order to enhance the system performance and minimize the bottlenecks.

Allocation Policies that are Energy Aware

Incorporate energy-saving in the process of resource allocation. Take into account the workload consolidation, effective use of equipment, and active adjustments in accordance with energy conscious policy to minimize operation expenses and environmental consequences.

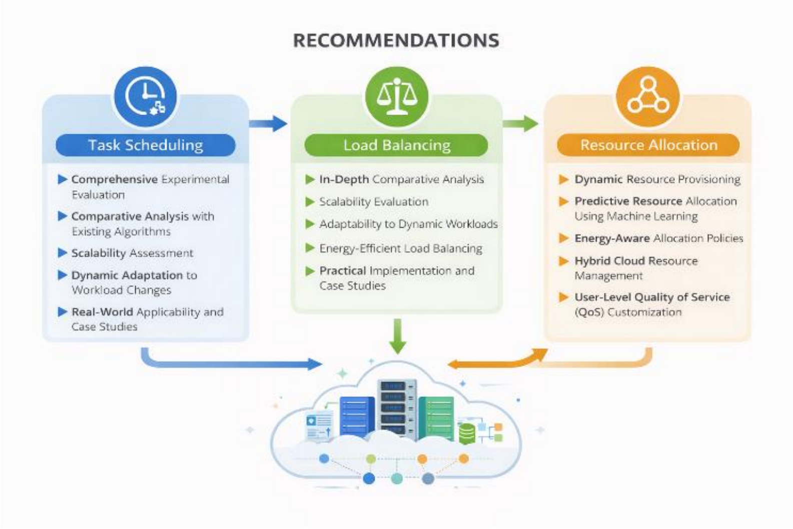
Cloud Resource Management Hybrid

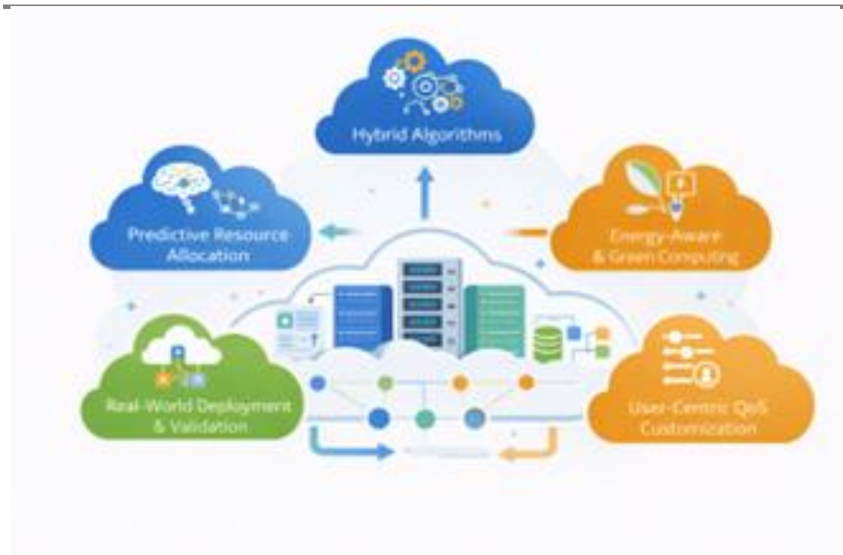
Use hybrid approaches, involving the use of both public and private cloud capabilities. Create smart allocation programs that assign workload according to cost, performance and data privacy needs and make them work optimally across the environments.

Quality of Service (QoS) Customization User-Level

Give the users the capability to specify QoS parameters, including performance priorities or latency sensitivity, or workload constraints. Scalable allocation gives users the power to customize cloud resources in line with application-specific requirements and enhance general satisfaction and system efficiency.

An overview of the suggested resource optimization strategies in cloud computing, where the focus is placed on Task Scheduling, Load Balancing, and Resource Allocation and the most important actionable recommendations for each of these categories are given.


RECOMMENDATIONS
Task Scheduling
Comprehensive Experimental Evaluation
Comparative Analysis with Existing Algorithms
Scalability Assessment
Dynamic Adaptation to Workload Changes
Real-World Applicability and Case Studies
Load Balancing
In-Depth Comparative Analysis
Scalability Evaluation
Adaptability to Dynamic Workloads
Energy-Efficient Load Balancing
Practical Implementation and Case Studies
Resource Allocation
Dynamic Resource Provisioning
Predictive Resource Allocation Using Machine Learning
Energy-Aware Allocation Policies
Hybrid Cloud Resource Management
User-Level Quality of Service (QoS) Customization


Hybrid Algorithms
Predictive Resource Allocation
Energy-Aware & Green Computing
Real-World Deployment & Validation
User-Centric QoS Customization

CONCLUSION

This systematic review is a critical synthesis of peer-reviewed articles (37) related to cloud resource optimization published between 2018 and early 2025. This study uses structured inclusion criterion, quality evaluation and taxonomy-based analysis to find patterns, trade-offs and research gaps unlike descriptive surveys. The results show that although metaheuristic and reinforcement learning models have high adaptability, scalability and computational scalability are significant issues in large scale cloud setups. Moreover, the vast majority of studies pay attention to the performance optimization, but a smaller number of works combine cost, energy efficiency, and QoS constraints together. The review shows the growing trend towards AI-hybrid models and multi-cloud orchestration frameworks. Nonetheless, validity and reproducibility of this in the real world are not yet proven by literature. The research offers a systematic source to the researchers and practitioners wishing to develop scalability, adaptable, and energy-sensitive cloud optimization systems.

Future Work

Although significant advancements have been made in cloud resource optimization, several research opportunities remain open for further exploration. The way forward in the future is to consider the design of hybrid task scheduling and load balancing algorithms that combine various optimization methods, including metaheuristics, machine learning, and reinforcement learning, to enhance flexibility and efficiency in dynamic and large-scale cloud settings. The possibility of studying predictive and proactive resource allocation strategies based on developed machine learning models to predict workload patterns, predict system bottlenecks, and optimize resource consumption without negatively affecting the quality of services is also significant. Also, further studies may be dedicated to the practical application and verification of these strategies in heterogeneous, multi-cloud environments in the future to determine their practical relevance and scalability in the changing network, computing, and storage conditions. The implementation of energy conscious and green computing policies is an important field to be considered as cloud data centers tend to find green and sustainable operations. In addition, increasing the user centric customization of parameters of quality of service (QoS) like resource priority-based allocation and dynamic SLA tuning will enable providers of cloud services to be better placed in fulfilling various application needs. In summary, studies in the future should strive to fill the gap between theory and practice to provide high-strength, scalable, and energy-efficient solutions that suit the dynamic requirements of the new cloud computing infrastructures.

REFERENCE

1. N. et al., “A Systematic Literature Review for Load Balancing and Task Scheduling in Cloud Computing,” Artificial Intelligence Review, 2024. https://doi.org/10.1007/s10462-024-10925-w
2. S. et al., “Dynamic Load Balancing in Cloud Computing: Optimized RL Based Clustering & Task Scheduling,” Processes, vol. 12, 2024. https://doi.org/10.3390/pr12030519

3. A. et al., "Resource Allocation with Efficient Task Scheduling Using Hierarchical Auto Associative Neural Networks in Cloud Computing," Expert Systems with Applications, 2024. https://doi.org/10.1016/j.eswa.2024.123554
4. B. et al., "A Systematic Literature Review on Task Allocation & Performance Management in Cloud Data Centers," Computers, Systems & Engineering, 2024. https://doi.org/10.32604/csse.2024.042690
5. C. et al., "Optimization Based Resource Scheduling Techniques in Cloud Computing: Review & Future Directions," Computers & Electrical Engineering, 2025. https://doi.org/10.1016/j.compeleceng.2025.110080
6. D. et al., "A Novel QoS Aware Task Scheduling Approach Using Modified Wombat Optimization Algorithm," Journal of Engineering & Applied Science, 2025. https://doi.org/10.1186/s44147-025-00628-6
7. E. et al., "Performance Analysis of Cloud Computing Task Scheduling with Metaheuristic Algorithms," Electronics, 2025. https://doi.org/10.3390/electronics14101988
8. F. et al., "Cost Modelling and Optimisation for Cloud: A Graph Based Approach," Journal of Cloud Computing, 2024. https://doi.org/10.1186/s13677-024-00709-6
9. G. et al., "LLM Based Cost Aware Task Scheduling for Cloud Computing Systems," Journal of Cloud Computing, 2025. https://doi.org/10.1186/s13677-025-00822-0
10. H. et al., "Optimized Task Scheduling in Fog Cloud Computing Using Hybrid Deep Learning and Metaheuristic Algorithms," Neural Processing Letters, 2026. https://doi.org/10.1007/s11063-025-11819-w
11. I. et al., "Systematic Review: Load Balancing in Cloud Computing by Using Metaheuristic Dynamic Algorithms," Intelligent Automation & Soft Computing, 2024. https://doi.org/10.32604/iasc.2024.050681
12. J. et al., "Machine Learning Based Cloud Resource Allocation Algorithms: A Comprehensive Comparative Review," arXiv, 2025. https://arxiv.org/abs/2511.11603
13. K. Luo, H. Huang, C. Zhang, and S. Guo, "Task Scheduling in Cloud Computing: A Survey," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 3, pp. 675–699, 2022. https://doi.org/10.1109/TPDS.2021.3117298
14. L. Wang, Z. Li, and M. Zhou, "Task Scheduling in Cloud Computing: A Survey," IEEE Access, vol. 8, pp. 83184–83216, 2020. https://doi.org/10.1109/ACCESS.2020.2996766
15. J. Yu, Z. Li, J. Liu, and M. Zhou, "Machine Learning Based Approach for Load Balancing with Resource Constraints in Cloud Computing," Applied Soft Computing, vol. 120, 108677, 2022. https://doi.org/10.1016/j.asoc.2022.108677
16. M. Malhotra, R. Chopra, and R. Sibal, "Load Balancing Techniques in Cloud Computing Environment: A Review," Computer Networks, vol. 236, 103768, 2023. https://doi.org/10.1016/j.comnet.2023.103768
17. S. Pepe, F. Khan, and S. Singh, "Low Time Complexity & Low Cost Particle Swarm Optimization for Cloud Task Scheduling and Load Balancing," Applied Intelligence, vol. 49, no. 9, pp. 3308–3330, 2019. https://doi.org/10.1007/s10489-018-1304-7
18. R. Buyya, M. Vecchiola, and S. T. Selvi, Mastering Cloud Computing, 2nd ed., Morgan Kaufmann, 2023. https://doi.org/10.1016/C2022-0-05810-0
19. T. Qiu, X. Zhang, and L. Mao, "Resource Scheduling in Cloud Computing with Modified Genetic Algorithm," IEEE Access, vol. 7, pp. 104954–104966, 2019. https://doi.org/10.1109/ACCESS.2019.2936712
20. A. Beloglazov and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," Future Generation Computer Systems, 2018. https://doi.org/10.1016/j.future.2018.04.054
21. M. Mishra and A. Sahoo, "Energy-Aware VM Allocation in Cloud Data Centers Using Reinforcement Learning," Journal of Cloud Computing, 2021. https://doi.org/10.1186/s13677-021-00275-8
22. P. Kumar, Y. Singh, and A. Sharma, "A Survey on Resource Allocation Strategies in Cloud Computing," IEEE Access, 2022. https://doi.org/10.1109/ACCESS.2022.3145632
23. M. Z. A. Bhuiyan, S. T. Selvi, and M. A. R. Sarkar, "Survey on Scheduling Algorithms in Cloud Computing," International Journal of Computer Applications, 2019. https://doi.org/10.5120/ijca2019918871
24. J. Liu, Y. Zhao, and H. Xu, "Load Balancing Algorithms in Cloud Computing: A Review," IEEE Access, vol. 8, pp. 21490–21508, 2020. https://doi.org/10.1109/ACCESS.2020.2975269

25. S. Seneviratne, J. S. Sahu, and D. Chatterjee, “Task Scheduling Strategy Based on Ant Colony Optimization in Cloud Computing Systems,” International Journal of Adaptive Control and Signal Processing, vol. 34, no. 4, pp. 1259–1278, 2020. https://doi.org/10.1002/acs.3269
26. R. Mondal and S. Bhowmik, “Meta Heuristic Based Resource Scheduling Techniques for Enhancing Performance in Cloud Datacenters,” Cluster Computing, vol. 23, pp. 1337–1351, 2020. https://doi.org/10.1007/s10586-019-03015-3
27. P. Patel, V. Kumar, and M. Singh, “Task Scheduling Algorithms and Their Performance Parameters in Cloud Computing,” Journal of Cloud Computing, vol. 9, no. 1, 5, 2020. https://doi.org/10.1186/s13677-020-0153-y
28. K. Bilal, H. Hussain, S. U. Khan, and M. F. Zhani, “Communication Efficient Resource Management and Scheduling in Cloud,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 824–859, 2019. https://doi.org/10.1109/COMST.2018.2868538
29. S. Abbas and R. Buyya, “Resource Provisioning in Cloud Computing: Review & Open Challenges,” Journal of Cluster Computing, vol. 23, pp. 545–564, 2020. https://doi.org/10.1007/s10586-019-03014-4
30. M. A. Alworafi, A. Dhari, S. A. Al-Hashmi, and A. B. Darem, “A Survey on Task Scheduling in Cloud Computing,” IEEE Access, vol. 6, pp. 13474–13489, 2018. https://doi.org/10.1109/ACCESS.2018.2803249
31. A. Belgacem, S. Mahmoudi, and M. Kihl, “Intelligent Multi Agent Reinforcement Learning Model for Resource Allocation in Cloud Computing,” Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 6, pp. 2391–2404, 2022. https://doi.org/10.1016/j.jksuci.2022.03.016
32. S. Saxena and A. K. Singh, “Auto Adaptive Learning Based Workload Forecasting in Dynamic Cloud Computing Environments,” International Journal of Computers and Applications, vol. 43, no. 4, pp. 456–471, 2023. https://doi.org/10.1080/1206212X.2022.2145789
33. H. Shukur, S. R. M. Zeebaree, R. R. Zebari, and O. M. Ahmed, “Cloud Computing Virtualization for Resource Allocation in Distributed Systems: Trends & Challenges,” Journal of Applied Science and Technology Trends, vol. 3, no. 2, pp. 89–102, 2023. https://doi.org/10.32604/jastt.2023.034210
34. Y. Li, M. Chen, and W. Wang, “Deep Reinforcement Learning for Joint Task Scheduling and Resource Allocation in Cloud-Edge Computing,” IEEE Access, vol. 7, pp. 150006–150017, 2019. https://doi.org/10.1109/ACCESS.2019.2947393
35. Q. Qi, X. Xu, and H. Jin, “Adaptive Resource Allocation in Cloud Computing Using Q-Learning,” Future Generation Computer Systems, vol. 93, pp. 887–898, 2019. https://doi.org/10.1016/j.future.2018.12.027
36. Y. Zhao, X. Liu, and Y. Zhang, “Deep Reinforcement Learning for Dynamic Resource Allocation in Cloud Computing,” Future Generation Computer Systems, vol. 99, pp. 709–719, 2019. https://doi.org/10.1016/j.future.2019.05.062
37. S. Wang, Z. Liu, and X. Xu, “A Multi-Objective Resource Allocation Model in Cloud Computing Using Artificial Intelligence,” Journal of Network and Computer Applications, vol. 136, pp. 13–25, 2019. https://doi.org/10.1016/j.jnca.2019.03.006