Journal of Global Research in Electronics and Communication (JGREC)
Volume 2 | Issue 11 | November 2026 | Pages 28–36
DOI: https://doi.org/xxxxxxxxxxxxx
e-ISSN: 2321-3175
Review Article OPEN ACCESS Available online: https://jgrec.info/index.php/jgrec/index

Artificial Intelligence for Cloud Resource Management: A Survey of Optimization Techniques

Dr. Nilesh Jain
Associate Professor
Department of Computer Sciences and Applications
Mandsaur University, Mandsaur
nileshjainmca@gmail.com

Corresponding Author: * Nilesh Jain (nileshjainmca@gmail.com)

Received: 15 November 2026 | **Revised:** 20 November 2026 | **Accepted:** 25 November 2026

***Abstract*—Cloud resource management is essential for efficiently allocating, monitoring, scheduling, and utilizing computing, storage, network, and virtualized resources in dynamic cloud environments. This survey reviews resource management approaches and optimization techniques with emphasis on performance, energy efficiency, cost reduction, service quality, and resource utilization. It examines major resource management methods, including energy-aware, SLA-aware, market-based, load-balanced, network load-aware, hybrid cloud, and mobile cloud resource management. Resource allocation and scheduling approaches, including virtual machine allocation, dynamic allocation, workload allocation, static scheduling, dynamic scheduling, and workflow-based scheduling, are also discussed. The survey further analyzes artificial intelligence techniques for cloud resource management, including deep learning, blockchain integration, fog computing, and edge-cloud systems. Optimization techniques such as static and dynamic provisioning, genetic algorithms, stochastic approaches, approximate dynamic programming, Benders decomposition, and sample average approximation are reviewed. AI-based optimization is examined for cloud resource allocation, energy systems, supply chains, environmental management, and real-time optimization. Finally, the survey highlights challenges associated with dynamic workloads, heterogeneous resources, scalability, energy consumption, SLA requirements, security, and real-time decision-making, providing directions for efficient and intelligent cloud resource management. The findings emphasize adaptive, scalable, and sustainable strategies for cloud infrastructures.**

***Keywords*—Cloud Resource Management, Artificial Intelligence, Resource Allocation, Resource Scheduling, Optimization Techniques, Load Balancing, Virtual Machines.**

I. INTRODUCTION

The increasing demand for computational resources requires flexible, scalable, and cost-effective computing environments capable of supporting diverse applications. Cloud computing addresses these requirements by providing on-demand access to computing resources over the Internet. Cloud resource management systems provide the facility of resource provisioning as per the demand of scientific application's execution. Cloud computing is defined as providing the facility for cloud users to use the resources over the Internet for the execution of their applications as per the requirement. The management of resources in cloud environment is an NP-hard problem. Thus, it is challenging. A low-quality resource management system can potentially result in a long processing time and high cost [1]. The main goal of cloud resource management systems is to interact with the underlying hardware infrastructure and control the provisioning and scheduling of the resources for the successful execution of the application as per the user's requirements. Resource scheduling has become an important core field of cloud resource management systems as it is responsible for orchestrating the resources to both cloud providers and cloud users [2]. A resource scheduling algorithm plays an important role in the process of managing the huge number of resources of cloud data centres in an efficient manner to achieve the target of quality requirements with the help of resource monitoring, by managing the decision of resource-job mapping process.

Artificial Intelligence (AI) and Cloud Computing have emerged as two of the most transformative technologies of the 21st century, each revolutionising the way businesses operate and innovate. The convergence of these technologies is ushering in a new era of intelligent, adaptive, and efficient computing solutions that promise to reshape industries across the board. AI, with its ability to mimic human intelligence, has made significant strides in recent years. It excels in tasks that once required human cognition, such as visual perception, speech recognition, decision-making, and language translation [3]. From autonomous vehicles to personalized healthcare, AI is driving innovation and efficiency in diverse sectors [4]. AI-driven optimization in cloud computing encompasses a broad spectrum of techniques designed to improve resource utilization, reduce latency, and enhance overall service quality [5]. These methods often employ machine learning models to predict workload patterns, dynamically allocate resources, and optimize scheduling, thus ensuring efficient cloud operations under varying conditions. Techniques such as neural networks, reinforcement learning, and fuzzy logic have been successfully implemented to optimize resource management and mitigate the inefficiencies

Copyright © 2026 The Author(s).
This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, adaptation, and reproduction in any medium, provided the original author(s) and the source are properly credited.

J. Glob. Res. Elec. Comm., 2 (11):28-36
DOI: https://doi.org/xxxxxxxxxxxxx

associated with traditional cloud management approaches [6]. Artificial intelligence techniques for efficient cloud resource management, focus on resource allocation, workload prediction, scheduling, load balancing, energy efficiency, and cost optimization.

A. Structure of the Paper

The structure of the paper: Section II presents the role and fundamentals of cloud resource management. Section III discusses AI-based resource management. Section IV covers optimization techniques and AI-based optimization. Section V reviews recent literature and research. Section VI concludes the survey and presents future research directions.

II. Role of Cloud Resource Management

Cloud resource management involves the efficient allocation, monitoring, and utilization of computing, storage, network, and virtualized resources. It aims to maintain service quality while minimizing operational cost, energy consumption, and resource wastage. Effective management includes resource provisioning, workload scheduling, load balancing, auto-scaling, and performance monitoring to support dynamic cloud workloads.

A. Cloud Computing Resource Types

In cloud computing, usually, all resources are virtualized and shared among multiple consumers. There are several challenges in resource management such as network load, energy efficiency, SLA violation, load balancing, profit maximization, etc. The resource management metrics have multiple challenges, for example, network loads and energy efficiency introduce SLA violations, and reduction in SLA violations enhances energy consumption and affects the profit. Resource management is a vital aspect of cloud computing for providing better performance and efficient utilization of underlying hardware [7]. This section discusses different types of resource management methods.

1) Energy-Aware Resource Management

The high energy consumption can result in an increase in the power expenses of the service providers as well as an increase in CO_2 emissions. Therefore, energy efficiency is a significant issue in cloud computing. To deal with this issue, workload consolidation is important [8]. The energy consumption can be reduced through workload consolidation on a smaller number of servers. The VMs on the servers with less workload can be migrated to the servers with higher workload, and the idle servers can be turned off. Based on the modified best fit decreasing, the VMs can be sorted in descending order according to the requirements of CPU.

2) Service Level Agreement (SLA)-Aware Resource Management

The SLA signed between the service provider and consumer contains information about the required service level, price, and penalty clause imposed if agreement is violated [9]. Hence, SLA-aware resource management is significant. To manage fluctuating workload and ensure SLA, capacity allocation algorithms can play an important role[10]. A workload predictor can be used for predicting future workload requirements. While the virtual machines (VMs) communicate, the response time is kept lower to avoid SLA violations. Minimizing the probability of under- or over-resource provisioning is another vital aspect.

3) Market-Based Resource Management

Cloud computing is a market-oriented paradigm. Thus, the service providers have the objective of increasing their profit. Auction-based resource allocation has been discussed, where the user bids for the advertised resources, and if he/she wins then only resources are assigned. Multi-layer resource management has been discussed, where the Software as a Service (SaaS)provider provides service to the consumers only if they use the services of IaaS providers [11]. The SaaS provider uses resources of the IaaS provider and provides to the consumer. The IaaS provider uses an optimal resource allocation method for enhancing own revenue.

4) Load-Balanced Resource Management

Load balancing is an important feature in a computing environment. Load balancing usually refers to the process of workload sharing among multiple resources. A dynamic load balancing approach has been proposed, where the workload is shifted from an overloaded server to a server that is less utilized [12]. If two machines are less utilized, the workload of one machine is shifted to the other, and then switched off. A load balancing method based on an artificial neural network has been discussed. This method uses backpropagation for equal load distribution among all servers. Each user's demand is predicted and resource allocation takes place accordingly. However, at any given time, the active servers depend on the users' needs at that particular time. Thus, active servers are minimized and energy consumption is reduced.

5) Network Load-Aware Resource Management

In cloud computing, the consumer uses the service via Internet. Hence, the network load is an important factor. The network load refers to the amount of traffic passing through the network at some specific time [13]. In cloud computing, the resource managers share information, VM placement and migration take place, inter-VM communication occurs, hence, network load becomes vital. High network load degrades performance because of the waiting time for VM placement and delay in inter-VM communication [14]. Thus, minimization of traffic amount is significant, and network load-aware resource management is required.

6) Hybrid Cloud Resource Management

The hybrid cloud is usually the integration of public and private clouds. In this case, the decision regarding the resource use of public cloud is significant. A rule-based resource manager has been discussed. The user requests are of two types: critical tasks with higher priority and secondary tasks with low priority [15]. For securing the critical tasks or data, private cloud is selected for hosting them. On the other hand, the secondary tasks or data can utilize public as well as private cloud resources. However, the public cloud is only used if the private cloud is fully exhausted. The requests for VM provisioning can be categorized as low, medium, and high priority classes.

7) Mobile Cloud Computing (MCC) Resource Management

With the huge increase in the number of smartphone users, MCC has become a significant research domain. In traditional MCC, an agent-client architecture, where the mobile devices use the resources available on the cloud. Nevertheless, mobile devices can share resources in a cooperation-based architecture as they have ample resources [16]. MCC can be considered as an integration of various domains, where each domain contains cloud resources [17]. To manage cloud resources and provide continuous service to users, the load can be shared across domains. To improve customer satisfaction

J. Glob. Res. Elec. Comm., 2 (11):28-36
DOI: https://doi.org/xxxxxxxxxxxx

and reduce service rejections, service request decision-making can be modelled as a semi-Markov decision process [18].

B. Resource Allocation and Scheduling

In light of the rapid growth and evolution of cloud computing, the efficient allocation of resources and the judicious scheduling of tasks have become critical to improve performance and reduce costs. As organizations increasingly rely on cloud infrastructures to meet their computing requirements, understanding the intricacies of resource management has become critical [19]. These processes are fundamental and vital for optimizing the utilization of cloud resources while simultaneously meeting user requirements and provider objectives.

1) Resource Allocation in Cloud Computing

Resource allocation is the process of assigning appropriate resources to the tasks required by the consumer to ensure efficient completion [20]. In cloud computing, it refers to designating a virtual machine that meets the properties defined by the consumers. consumer submit the request to the resource allocator, the request added to the queue list, resource allocator informs the allocation unit about the request, allocation unit asks for the requested resources from the Infrastructure as a Service (IaaS),if the resources are available, figure 1 then IaaS respond back positively, the allocation unit creates a VM from VM pool according to the request, resource allocator is informed after creating a VM, requests are de-queued, and resources are allocated. The major resource allocation approaches are as follows:

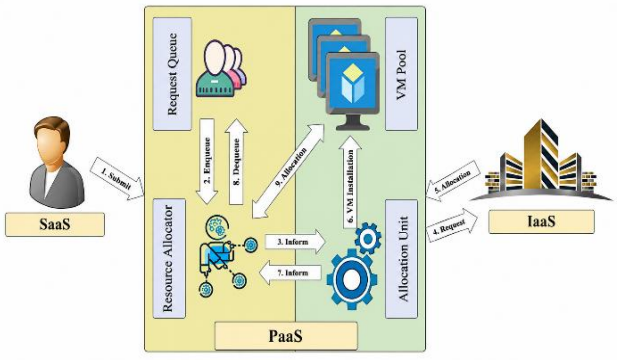


Fig. 1. The Basic Flow of Resource Allocation in Cloud Computing

- **Virtual Machine Allocation:** Virtualization technology is the main reason that makes cloud computing affordable and practical [21]. Cloud computing resources require high investments as they are very costly, so the cloud service provider wants to adopt techniques that allocate the virtual machine to the task efficiently.
- **Dynamic Resource Allocation:** The workload of the task submitted to the cloud infrastructure by the consumers changes continuously [22]. In order to address the individual requirements of each task, cloud service provider has to use dynamic techniques to allocate resources.
- **Workload Resource Allocation:** Workload generally reflects the ability of the system to handle and process the task. Workload should be sufficient on a system to do the tasks efficiently within the cloud environment. This parameter determines the amount of workload on empirical setup of resource allocation techniques.

2) Resource Scheduling in cloud computing

Resource scheduling is critical research area in the cloud environment because of the substantial resource cost and execution time. Various resource parameters and scheduling criteria are considered in various categories of resource scheduling techniques [23]. Resource scheduling becomes more difficult because both cloud resource provider and cloud service consumer do not want to share their data with each other. The major resource scheduling approaches are as follows:

- **Static Scheduling:** Static scheduling is based on the idea that efficient hardware utilization can be achieved when operations begin to share resources without sacrificing performance. When the code contains operations with data and control dependencies that are statically indeterminable or have variable latency, static scheduling plans the beginning times of specific operations to accommodate the worst-case timing scenario.
- **Dynamic Scheduling:** Dynamic scheduling of operations is possible during runtime. By getting beyond the rigidity of conventional static scheduling, it improves throughput in control-dominated and irregular applications. Additional applications that dynamic scheduling can manage include those whose memory accesses are unknown while the application is being built.
- **Workflow-Based Scheduling:** The applications in workflow scheduling are depicted as directed acyclic graphs (DAGs), as illustrated [24]. This makes it a generic form of the task scheduling problem. There is a tendency for the terms workflow and DAG to be used interchangeably.

III. Artificial Intelligence for Cloud Resource Management

Artificial Intelligence (AI) enhances cloud resource management by enabling intelligent prediction, allocation, scheduling, and optimization of computing resources. Machine learning, deep learning, and reinforcement learning can analyze workload patterns and dynamically adjust resources to improve performance, reduce costs, minimize energy consumption, and maintain efficient utilization in highly dynamic cloud environments.

A. Deep Learning and the Blockchain for Cloud Resource Management

Blockchain integrated with deep learning presents a paradigm shift in the management of cloud resources. The synergy helps solve critical challenges of trust, security, and decentralized decision-making to present a robust and efficient framework for DRA [25]. The major applications of these technologies include:

- **Decentralization and Trust:** Blockchain inherently provides a decentralized framework that can instil trust among participants in the cloud ecosystem [26]. When combined with DL, resource allocation decisions become transparent and immutable, ensuring accountability and reducing reliance on centralized entities.
- **Improved Security:** The cryptographic features of blockchain ensure data integrity and prevent unauthorized access [27]. When integrated with DL

J. Glob. Res. Elec. Comm., 2 (11):28-36
DOI: https://doi.org/xxxxxxxxxxxxx

models, blockchain secures both the training data and resource allocation processes, mitigating risks such as data tampering and adversarial attacks.

- **Automate Resource Allocation:** Smart Contracts Blockchain smart contracts automatically implement resource allocation policy. These are powered by the predictions from DL, which dynamically allocate resources based on predefined criteria for better operational efficiency and less human intervention.
- **Auditability and Compliance:** The immutable ledger of blockchain allows for detailed tracking of resource allocation events to ensure that all regulatory standards are met [28]. DL algorithms can analyze these logs for anomaly detection or future optimization strategies.
- **Resource Tokenization:** Blockchain allows the tokenization of cloud resources, which enables dynamic pricing and efficient resource sharing [29]. DL models predict demand and optimize token distribution, ensuring equitable resource access and cost-effectiveness.
- **Energy Efficiency:** Combining DL's predictive capabilities with blockchain's decentralized consensus mechanisms optimizes energy consumption [30]. For example, DL models can predict peak demand periods, while blockchain-based smart grids allocate energy resources accordingly, minimizing waste.
- **Integration with Federated Learning:** Blockchain enables the integration of federated learning models by securely sharing data, collaborative training while preserving data privacy, and efficiently managing distributed resources [31].

B. *Artificial Intelligence's of Resource Management*

Artificial intelligence (AI) technology has emerged as a powerful tool to improve resource management efficiency in cloud computing systems. AI includes various sub-fields such as machine learning, reinforcement learning, predictive analytics, natural language processing, and genetic algorithms[32]. These technologies enable intelligent decision-making, automation, and optimization based on data analysis, pattern recognition, and learning from historical and real-time data. Using AI, cloud computing systems can dynamically allocate resources, optimize workload scheduling, predict resource requirements, detect anomalies and errors, and perform energy-efficient operations. AI-based resource management enables proactive and adaptive strategies that improve the efficiency, scalability, performance, and cost-effectiveness of cloud computing environments.

1) *Fog Computing Systems*

Fog computing is a distributed computing infrastructure where resource-intensive functions reside between the cloud and data sources, reducing the strain on resources (computing power, network bandwidth, and storage capacity) [33]. It's structured.

2) *Edge-Cloud Systems*

In edge cloud systems, resource functions such as computing power, networking, and storage are distributed throughout the system and closer to the traffic source. Figure 2 shows an edge cloud architecture with mobile devices [34]. The diagram assumes that the mobile device's edge cloud is linked to edge cloud server A. Mobile device (sub-service) tasks from the central cloud server are copied to edge cloud server A. At this point, when the mobile device moves to another location closest to Edge Cloud Server B, the related tasks (sub-services) are scheduled to migrate from Edge Cloud Server A to B.

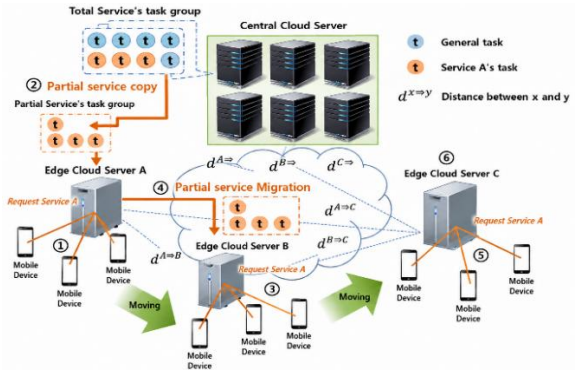


Fig. 2. The Edge-Cloud Architecture with Mobile Devices

IV. Cloud Resource Management of Optimization Techniques

Cloud resource management optimization techniques aim to improve the utilization, performance, and efficiency of distributed cloud resources while reducing operational costs and energy consumption. Heuristic, metaheuristic, evolutionary, swarm-intelligence, fuzzy-logic, and hybrid optimization methods are widely applied to address complex resource-management problems. These techniques support efficient decision-making under dynamic workloads and heterogeneous cloud environments.

A. *Optimization techniques*

In Computer Science and operation research, Optimization is referred to as the selection of best element from a set of alternatives with regard to some criteria. Two types of entities involved in cloud are cloud service providers and cloud consumers. Cloud service providers provision their resources on rental basis to cloud consumers and cloud consumers submit their requests for provisioning the resources [35]. They both have their own motivations when they become part of cloud environment.

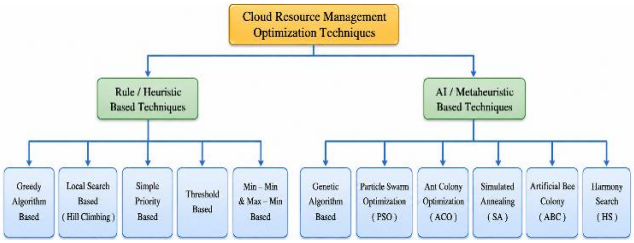


Fig. 3. Optimization Techniques

Figure 3 Consumers are concerned with the performance of their applications, whereas providers are more interested in efficient utilization of their resources.

1) *Static Provisioning*

The cloud service provider allocates resources to the consumer in advance. That is cloud consumer has to decide how much capacity of resources he/she wants statically, meaning before using them. The deterministic approach, integer programming, linear programming, and deadline provisioning algorithm show how resources are allocated statically for deadline-based workflows. Here consider only

DOI: https://doi.org/xxxxxxxxxxxx

the expected values for all the parameters. In static provisioning may be an over-provisioning or under-provisioning problem.

2) Dynamic Provisioning

The cloud service provider allocates resources to consumer as needed and when required and the consumer has to pay per use. This is also called the pay-as-you-go model. Here, genetic algorithm, stochastic approach, approximate dynamic programming, Benders decomposition and average approximation are used for provisioning resources dynamically. The major technologies of these include:

- **Genetic algorithm:** Genetic algorithms are pay commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, cross-over and selection.
- **Stochastic approach:** In this section, the stochastic programming with multistage recourse is presented as the core formulation. First, the original form of stochastic integer programming formulation is derived.
- **Approximate dynamic programming:** p-roximate Dynamic Programming (ADP) is a powerful technique to solve large-scale discrete-time multistage stochastic control processes, i.e., complex Markov Decision Processes (MDPs).
- **Benders decomposition:** The goal of this algorithm is to break down the optimization problem into multiple smaller problems which can be solved independently and parallel. The Benders decomposition algorithm can decompose integer programming problems with complicating variables into two major problems.
- **Sample average approximation:** It may not be efficient to get the solution of the algorithm if the number of scenarios is large by solving the stochastic programming formulation defined directly if all scenarios in the problem are considered.

B. AI-based resource optimization techniques

AI-based resource optimization techniques improve cloud resource utilization by making intelligent decisions about allocation, provisioning, scheduling, and workload distribution. Machine learning, deep learning, reinforcement learning, and metaheuristic methods analyze resource demands and system conditions to optimize performance, energy consumption, operational cost, and quality of service in dynamic cloud environments.

1) Manufacturing Resource Optimization

Many practitioners in the manufacturing industry are embracing AI technologies to optimize resource utilization to enhance productivity [36]. These efforts are a part of larger trends of lean production, just-in-time production and the advent of the Fourth Industrial Revolution (Industry 4.0). In the past few years, several AI solutions for scheduling, production scheduling and dynamic resource allocation have shown enormous gains in the use of machines across multiple production scheduling layers.

2) Energy System Management

Energy resource optimization is one of the best examples of AI applications, given the importance of energy efficiency and transition to renewable energy supply [37]. Smart grid technologies employ AI algorithms to optimize energy delivery, load balancing, and demand response programs. In addition, machine learning models can predict energy demand patterns with high accuracy, so utilities can effectively optimize generation resources and minimize operating expenses.

3) Supply Chain Optimization

AI applications have transformed supply chain resource optimization to address the challenges of a global supply network[38]. Machine learning algorithms that analyze historical demand data, seasonality data, and external impacts to improve inventory levels across the supply chain, have reduced inventory carrying costs between 15-25% while improving service level [39].

4) Cloud Computing Resource Allocation

The challenges of optimizing resources in cloud computing environments are different from traditional platforms due to their dynamic nature, heterogeneous resources, and variable workloads [40]. AI algorithms have emerged as critical methods for managing virtualized resources while adhering to service level agreements (SLAs). Machine learning models utilize existing demand patterns to predict resource requirements and automatically scale resources to reduce costs and enhance performance.

5) Environmental Resource Management

Environmental resource management has increasingly employed AI methods to solve sustainability issues and improve resource harvesting. Water resource management systems utilize machine learning algorithms to provide optimized water distribution, demand prediction, and leak detection in distribution investments. These systems have shown to provide reductions in water loss of between 20-30% in urban distribution systems.

6) Real-Time Optimization Systems

Making decisions on resource optimization in real-time necessitates the use of algorithms that can perform within effective timeframes and quality of solutions. Of particular interest are streaming machine learning algorithms whose main feature is the ability to incrementally update an optimization model while processing a continuous stream of data. The importance of these systems is highlighted in examples such as traffic management and energy grid management, where there are costs for delays and missed opportunities.

V. Literature Review

Recent literature highlights the use of artificial intelligence in cloud resource management to improve resource utilization and operational efficiency. AI-based approaches support dynamic resource allocation, provisioning, scheduling, load balancing, and optimization, enabling cloud infrastructures to respond effectively to changing workloads while reducing energy consumption, cost, and resource wastage.

P. N. Sutar and R. Gunjan (2026) present an AI-Based multi-objective resource scheduling framework that aims to optimize cost and performance in dynamic cloud settings. The proposed framework integrates cost-sensitive resource scheduling, multidimensional resource optimization for compute, storage, and network resources, workload forecasting using LSTM, and anomaly detection with Isolation Forest. The proposed framework can adapt to varying workloads and pricing schemes (on-demand, reserved, and spot instances) and supports ETL workflows,

DOI: https://doi.org/xxxxxxxxxxxxx

real-time analytics, microservices, and AI computing. Synthetic workload traces and real-world cloud traces are employed in the experimental evaluation to demonstrate the efficacy of cost savings of up to 30-40% while maintaining SLA compliance and improving system responsiveness [41].

R. Cigiri, et al. (2026) identify the possibility of smart automation in cloud resource management and give a basis for how to advance self-optimizing cloud infrastructure. The future project suggests a machine learning-based approach towards making optimal use of resources on the basis of predictive analytics and adaptive scheduling methods. Having been developed as a part of the project, the model combines algorithms of supervised learning, such as Random Forest and Gradient Boosting, to predict workload trends and allocate virtualized resources based on those trends. Simulated cloud data with different levels of workloads was tested experimentally [42].

F. Asani, et al. (2025) introduced Agentic AI, a novel approach that integrates autonomous, goal-driven, and collaborative agents to optimize cloud resource allocation. Through a systematic literature review, the study highlights key gaps in current techniques and proposes a conceptual framework based on Agentic AI principles. The framework enables agents to predict workloads, set system-wide goals, and dynamically allocate resources in real time. By bridging existing gaps, this research offers a unified and adaptable solution to improve scalability, workload distribution, and cost efficiency, advancing intelligent cloud resource management [43].

R. David, et al. (2025) used cloud resources to help increase operational efficiency, cost reduction through improvement in cloud resource allocation, dynamic scaling, performance optimization using AI and ML. The current paper explores multiple approaches utilizing artificial intelligence, including RL, deep learning, and data analytics to analyze, predict resource demand and to actively manage cloud resources, adaptively to demand in real time. Such techniques enable proactive resource allocation, predictive scalability based on usage trends, and minimization of latency, while maximizing computational power. In addition, ML models are used to predict forthcoming workloads, enabling resource allocation in terms of storage, processing and network bandwidth, making cloud systems more cost-effective and responsive [44].

Y. L. M. Latha, et al. (2024) provided an overview of AI-based approaches for effective resource management in cloud computing services, applications, and systems. This study aims to provide a comprehensive analysis of the ways in which artificial intelligence and cloud computing are impacting education, with a focus on accessibility, effectiveness, and the calibre of the educational process. Found that artificial intelligence led to a 25% improvement in the personalization of educational information, and cloud computing led to a 60% rise in the capacity of simultaneous users [45].

J. Barach and J. H. Muzamal (2024) provided an in-depth analysis of optimizing resource allocation in multi-cloud environments through the integration of advanced machine learning and optimization techniques. implemented a comprehensive strategy that combines CNN for predicting resource demands, Ant Colony Optimization (ACO) for optimizing initial resource allocation, PPO for dynamic adjustments, and GBM for forecasting energy consumption. The CNN model exhibited strong performance in demand prediction, with detailed metrics demonstrating its effectiveness across various scenarios. ACO significantly improved resource distribution efficiency, as evidenced by a reduction in active cloud regions [46].

S. Nagasundaram, et al. (2023) proposed an intelligent resource management method that controls mobile resources by monitoring device states and projecting their future stability. It looks at the possibility of using suggested resource management system on a variety of cloud-based platforms Artificial Intelligence (AI) approaches, such as big data, mobile computing, and Internet of Things (IoT) have become more common in the computer science area in recent years. It is critical to manage resources effectively to maintain service quality, service-level agreements, and the system's overall availability. AI may be used to address the needs of cloud resource management [47].

Although recent studies demonstrate the potential of artificial intelligence, machine learning, predictive analytics, and optimization techniques for cloud resource management, several research gaps remain. Existing approaches mainly focus on specific objectives such as cost reduction, workload prediction, resource allocation, energy optimization, or SLA maintenance rather than a unified intelligent framework. Furthermore, reliance on simulated workloads, synthetic traces, or limited cloud environments restricts applicability to complex and heterogeneous real-world infrastructures. The integration of compute, storage, network, energy, and workload characteristics remains insufficiently explored. Agentic AI offers promising autonomous capabilities, but practical validation, explainability, security, and governance require further investigation. Multi-cloud resource management also faces challenges related to interoperability, scalability, dynamic pricing, heterogeneous platforms, and real-time decision-making. Additionally, continuous adaptation to unpredictable workload changes remains limited. Table I summarizes the major research gaps identified from the reviewed studies. These gaps highlight the need for scalable, adaptive, and intelligent frameworks integrating predictive analytics, autonomous decision-making, optimization, and real-time monitoring for efficient and reliable resource utilization.

TABLE I. SUMMARY OF KEY STUDIES ON ARTIFICIAL INTELLIGENCE FOR CLOUD RESOURCE MANAGEMENT

References	Study Focus	Methodology / Technology	Key Findings	Challenges / Limitations	Future Direction
P. N. Sutar and R. Gunjan (2026)	AI-Based Multi-Objective Resource Scheduling	LSTM, Isolation Forest, cost-sensitive scheduling, multidimensional resource optimization	Achieved up to 30–40% cost savings while maintaining SLA compliance and improving responsiveness	Evaluation uses synthetic and selected real-world cloud traces	Develop adaptive scheduling for larger and more diverse cloud workloads
R. Cigiri et al. (2026)	Smart Cloud Resource Automation	Random Forest, Gradient Boosting, predictive analytics, adaptive scheduling	Predicted workload trends and allocated virtualized resources according to changing workloads	Evaluation based on simulated cloud data	Validate intelligent scheduling using real-time and heterogeneous cloud workloads

F. Asani et al. (2025)	Agentic AI for Cloud Resource Allocation	Agentic AI, autonomous agents, predictive workload analysis, dynamic allocation	Supports autonomous resource allocation, scalability, workload distribution, and cost efficiency	Proposed framework requires practical implementation and validation	Develop collaborative Agentic AI systems for real-time cloud optimization
R. David et al. (2025)	AI-Driven Cloud Resource Optimization	Reinforcement Learning, deep learning, data analytics, predictive scaling	Enables proactive resource allocation, reduced latency, and improved cost-effectiveness	Multiple AI approaches require effective workload and resource data	Integrate adaptive AI models for continuous resource optimization
Y. L. M. Latha et al. (2024)	AI-Based Cloud Resource Management	AI, cloud computing, mixed-research methodology	Reported improved scalability, personalization, and reduced administrative errors	Focus includes educational cloud applications, limiting broader generalization	Extend AI-based resource management to diverse cloud environments
J. Barach and J. H. Muzamal (2024)	Multi-Cloud Resource Allocation	CNN, Ant Colony Optimization, PPO, Gradient Boosting	Improved demand prediction, resource distribution, dynamic allocation, and energy forecasting	Multi-model optimization increases system complexity	Develop scalable and efficient AI optimization for heterogeneous multi-cloud systems
S. Nagasundaram et al. (2023)	Intelligent Cloud Resource Management	AI, big data, mobile computing, IoT, resource-state monitoring	Supports resource management, service quality, SLA compliance, and system availability	Broad conceptual application with limited details on large-scale deployment	Develop predictive and adaptive AI mechanisms for real-time cloud resource management

VI. Conclusion and Future Work

Cloud resource management is a fundamental component of cloud computing for achieving efficient utilization, reliable performance, and cost-effective service delivery. This survey reviewed major resource management approaches, including energy-aware, SLA-aware, market-based, load-balanced, network load-aware, hybrid, and mobile cloud resource management. Resource allocation and scheduling techniques were examined to understand their role in assigning resources, managing workloads, and improving execution efficiency. The study also discussed artificial intelligence for cloud resource management, including deep learning, blockchain integration, fog computing, and edge-cloud systems. Various optimization techniques, such as static and dynamic provisioning, genetic algorithms, stochastic approaches, approximate dynamic programming, Benders decomposition, and sample average approximation, were analyzed. AI-based optimization further supports adaptive resource allocation, workload prediction, energy management, and real-time decision-making. Overall, effective cloud resource management requires adaptive and scalable strategies capable of handling dynamic workloads, heterogeneous resources, energy constraints, SLA requirements, and changing performance demands while maintaining efficient and sustainable cloud operations.

Future research should focus on integrating AI, blockchain, and advanced optimization techniques for autonomous cloud resource management. Developing scalable real-time models for dynamic workloads, energy efficiency, SLA compliance, security, and heterogeneous resources can improve adaptive decision-making and support sustainable cloud infrastructures.

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of interest.

How to cite this article:

N. Jain, "Artificial Intelligence for Cloud Resource Management: A Survey of Optimization Techniques," *Journal of Global Research in Electronics and Communication*, vol. 2, no. 11, pp. 28–36, Nov. 2026. Doi: XXXX

References

[1] R. Aron and A. Abraham, "Resource scheduling methods for cloud computing environment: The role of meta-heuristics and artificial intelligence," *Eng. Appl. Artif. Intell.*, vol. 116, p. 105345, Nov. 2022, doi: 10.1016/j.engappai.2022.105345.

[2] F. Xhafa and A. Abraham, "Computational models and heuristic methods for Grid scheduling problems," *Futur. Gener. Comput. Syst.*, vol. 26, no. 4, pp. 608–621, Apr. 2010, doi: 10.1016/j.future.2009.11.005.

[3] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, Dec. 2016, doi: 10.1186/s40537-016-0059-y.

[4] M. Mittal, "Cloud Computing In Healthcare: Transforming Patient Care And Operations," *Int. J. Comput. Eng. Technol.*, vol. 15, no. 6, pp. 1920–1929, Dec. 2024, doi: 10.34218/IJCET_15_06_164.

[5] S. Ratnayake, "A Comprehensive Review Of Ai-Driven Optimization, Resource Management, And Security In Cloud Computing Environments," *Int. J. Sustain. Infrastruct. Cities Soc.*, vol. 9, no. 5, 2024.

[6] Y. Jani, "Efficiency and Efficacy: AWS Instance Benchmarking of Stable Diffusion 1.4 for AI Image Generation," vol. 4, no. 2, 2023.

[7] S. Mustafa, B. Nazir, A. Hayat, A. U. R. Khan, and S. A. Madani, "Resource management in cloud computing: Taxonomy, prospects, and challenges," *Comput. Electr. Eng.*, vol. 47, pp. 186–203, 2015, doi: 10.1016/j.compeleceng.2015.07.021.

[8] A. Mukherjee, D. De, and R. Buyya, "Cloud Computing Resource Management," in *Studies in Big Data*, vol. 151, Springer Nature Singapore, 2024, pp. 17–37. doi: 10.1007/978-981-97-2644-8_2.

[9] S. Kukar, "Leveraging Price-Quote Data with AI for Smarter Configure-Price-Quote Systems," in *2026 2nd International Conference on Artificial Intelligence Systems (AIS)*, IEEE, May 2026, pp. 329–335. doi: 10.1109/AIS69919.2026.11621526.

[10] D. Ardagna, S. Casolari, M. Colajanni, and B. Panicucci, "Dual time-scale distributed capacity allocation and load redirect algorithms for cloud systems," *J. Parallel Distrib. Comput.*, vol. 72, no. 6, pp. 796–808, Jun. 2012, doi: 10.1016/j.jpdc.2012.02.014.

[11] S. Zaman and D. Grosu, "A Combinatorial Auction-Based Mechanism for Dynamic VM Provisioning and Allocation in Clouds," *IEEE Trans. Cloud Comput.*, vol. 1, no. 2, pp. 129–141,

2013, doi: 10.1109/TCC.2013.9.

[12] Y. Ban, H. Chen, and Z. Wang, "EALARM: Enhanced Autonomic Load-Aware Resource Management for P2P Key-Value Storage in Cloud," in *2013 IEEE Seventh International Symposium on Service-Oriented System Engineering*, IEEE, Mar. 2013, pp. 150–155. doi: 10.1109/SOSE.2013.38.

[13] Y. Muni, "Advanced Multi-Protocol Label Switching (MPLS)-Enabled Networks: A Survey of Emerging Approaches Via Traffic Engineering," *Eng. Technol. J.*, vol. 11, no. 01, Jan. 2026, doi: 10.47191/etj/v11i01.22.

[14] G. Jung and K. M. Sim, "Agent-Based Adaptive Resource Allocation on the Cloud Computing Environment," in *2011 40th International Conference on Parallel Processing Workshops*, IEEE, Sep. 2011, pp. 345–351. doi: 10.1109/ICPPW.2011.18.

[15] S. Malik, F. Huet, and D. Caromel, "Reliability-aware scheduling in cloud computing," 2012, pp. 194–200.

[16] R. Lingam, "AI-Powered Framework for Evaluating Child-Friendly Mobile Applications," *Int. J. Hum. Comput. Intell.*, vol. 4, no. 4, pp. 535–549, 2025, doi: 10.5281/zenodo.15624307.

[17] G. Fei, J. Niu, and Z. He, "A Research on Mobile Cloud Computing and Future Trends," *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, vol. 3, p. 151159, 2016, doi: 10.4108/eai.21-4-2016.151159.

[18] R. Palwe, "The Role of User Experience Design in Advancing Digital Sustainability," *Int. J. Sci. Res.*, vol. 14, no. 10, pp. 1168–1176, 2025.

[19] S. Shiekh, M. Shahid, M. Sambare, R. A. Haidri, and D. K. Yadav, "A load-balanced hybrid heuristic for allocation of batch of tasks in cloud computing environment," *Int. J. Pervasive Comput. Commun.*, vol. 19, no. 5, pp. 756–781, Nov. 2023, doi: 10.1108/IJPCC-06-2022-0220.

[20] M. F. Manzoor, A. Abid, M. S. Farooq, N. A. Nawaz, and U. Farooq, "Resource Allocation Techniques in Cloud Computing: A Review and Future Directions," *Elektron. ir Elektrotechnika*, vol. 26, no. 6, pp. 40–51, Dec. 2020, doi: 10.5755/j01.eie.26.6.25865.

[21] T. Selvi Somasundaram *et al.*, "CARE Resource Broker: A framework for scheduling and supporting virtual resource management," *Futur. Gener. Comput. Syst.*, vol. 26, no. 3, pp. 337–347, Mar. 2010, doi: 10.1016/j.future.2009.10.005.

[22] A. T. Saraswathi, Y. R. A. Kalaashri, and S. Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing," *Procedia Comput. Sci.*, vol. 47, pp. 30–36, 2015, doi: 10.1016/j.procs.2015.03.180.

[23] G. Siva Nageswara Rao, N. Srinivasu, S. V. N. Srinivasu, and G. R. Koteswara Rao, "Dynamic Time Slice Calculation for Round Robin Process Scheduling Using NOC," *Int. J. Electr. Comput. Eng.*, vol. 5, no. 6, p. 1480, Dec. 2015, doi: 10.11591/ijece.v5i6.pp1480-1485.

[24] F. Wu, Q. Wu, and Y. Tan, "Workflow scheduling in cloud: a survey," *J. Supercomput.*, vol. 71, no. 9, pp. 3373–3418, 2015, doi: 10.1007/s11227-015-1438-4.

[25] A. Zharkin, V. Novskyi, V. Popov, and S. Palachov, "Improving the Reliability and Power Quality in Distribution Networks with Sources of Dispersed Generation," 2022, pp. 23–45. doi: 10.1007/978-3-030-82926-1_2.

[26] M. Selvi and B. Ramakrishnan, "Secured message broadcasting in VANET using Oppositional Flower Pollination Algorithm (OFPA)," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, Feb. 2021, pp. 1–6. doi: 10.1109/ICICV50876.2021.9388394.

[27] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modelling Long- and Short-Term Temporal Patterns with Deep Neural Networks," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, in SIGIR '18. New York, NY, USA, NY, USA: ACM, Jun. 2018, pp. 95–104. doi: 10.1145/3209978.3210006.

[28] R. K. Kanneganti, "Security and Compliance Issues in Cloud-Based Deployments of Content and Workflow Management Systems," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 01, pp. 131–138, Aug. 2024, doi: 10.58812/esiscs.v2i01.1122.

[29] W. Shi, H. Matsumiya, H. Nishikawa, X. Kong, and H. Tomiyama, "Fusing Infrared and Visible Images for DNN-based Nighttime Human Detection," in *2022 19th International SoC Design Conference (ISOCC)*, IEEE, Oct. 2022, pp. 203–204. doi: 10.1109/ISOCC56007.2022.10031459.

[30] Y. M. Ismail, D. H. Hussein, S. Askar, and M. A. Ibrahim, "Deep Learning Techniques for Network Security," *Indones. J. Comput. Sci.*, vol. 14, no. 1, Feb. 2025, doi: 10.33022/ijcs.v14i1.4737.

[31] D. H. Abdulazeez and S. K. Askar, "A Novel Offloading Mechanism Leveraging Fuzzy Logic and Deep Reinforcement Learning to Improve IoT Application Performance in a Three-Layer Architecture Within the Fog-Cloud Environment," *IEEE Access*, vol. 12, pp. 39936–39952, 2024, doi: 10.1109/ACCESS.2024.3376670.

[32] S. Kanungo, "AI-driven resource management strategies for cloud computing systems, services, and applications," *World J. Adv. Eng. Technol. Sci.*, vol. 11, no. 2, pp. 559–566, Apr. 2024, doi: 10.30574/wjaets.2024.11.2.0137.

[33] A. A. Mutlag, M. K. Abd Ghani, N. Arunkumar, M. A. Mohammed, and O. Mohd, "Enabling technologies for fog computing in healthcare IoT systems," *Futur. Gener. Comput. Syst.*, vol. 90, pp. 62–78, Jan. 2019, doi: 10.1016/j.future.2018.07.049.

[34] Q. Cui *et al.*, "Stochastic Online Learning for Mobile Edge Computing: Learning from Changes," *Comm. Mag.*, vol. 57, no. 3, pp. 63–69, Mar. 2019, doi: 10.1109/MCOM.2019.1800644.

[35] K. Sumalatha and M. S. Anbarasi, "A review on various optimization techniques of resource provisioning in cloud computing," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, p. 629, Feb. 2019, doi: 10.11591/ijece.v9i1.pp629-634.

[36] F. Sheikh, M. Gupta, R. Ahluwalia, N. Gupta, and T. Chugh, "Artificial intelligence in dentistry: A comprehensive review of past, present, and future directions," *Santosh Univ. J. Heal. Sci.*, vol. 11, no. 2, pp. 324–326, Jul. 2025, doi: 10.4103/sujhs.sujhs_108_25.

[37] Z. T. Nayyef, M. M. Abdulrahman, and N. Al. Kurdi, "Optimizing Energy Efficiency in Smart Grids Using Machine Learning Algorithms: A Case Study in Electrical Engineering," *SHIFRA*, vol. 2024, pp. 46–54, Apr. 2024, doi: 10.70470/SHIFRA/2024/006.

[38] S. Kadyrov, A. Azamov, Y. Abdumajitov, and C. Turan, "Deep reinforcement learning for dynamic vehicle routing with demand and traffic uncertainty," *Oper. Res. Perspect.*, vol. 15, p. 100351, Dec. 2025, doi: 10.1016/j.orp.2025.100351.

[39] S. Remella *et al.*, "Intelligent Cost Governance in Cloud Supply Chains," in *2025 1st International Conference on Advancement in Futuristic Technologies (ICAFT)*, Belagavi, India: IEEE, 2025, pp. 1–7, December. doi: 10.1109/ICAFT66710.2025.11452715.

[40] P. Nawrocki and P. Osypanka, "Cloud Resource Demand Prediction using Machine Learning in the Context of QoS Parameters," *J. Grid Comput.*, vol. 19, no. 2, p. 20, Jun. 2021, doi: 10.1007/s10723-021-09561-3.

[41] P. N. Sutar and R. Gunjan, "AI-Driven Multi-Objective Resource Scheduling for Cost and Performance Optimization in Dynamic Cloud Computing Environments," in *2026 2nd International Conference on Computing, Communication and Green Engineering (CCGE)*, IEEE, May 2026, pp. 1–5. doi: 10.1109/CCGE67142.2026.11581845.

[42] R. Cigiri, N. Suma, V. Akshitha, S. Pravalika, V. Hasini, and N. Sheershika, "Intelligent Resource Optimization in Cloud Computing Using Machine Learning Techniques," in *2026 International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, Jun. 2026, pp. 1–6. doi: 10.1109/ICSCAN66520.2026.11588341.

[43] F. Asani, J. Ajdari, and X. Zenuni, "Agentic AI for Dynamic Cloud Resource Management: A Conceptual Framework for Dynamic Cloud Resource Management," in *2025 MIPRO 48th ICT and Electronics Convention*, IEEE, Jun. 2025, pp. 949–954. doi: 10.1109/MIPRO65660.2025.11131772.

[44] R. David, D. K. Arivalagan, A. S. Prabu, S. Ravi, P. Giri, and A. A. A. Samhan, "Optimizing Cloud Computing Efficiency for Scalable and Adaptive Resource Management via AI and Machine Learning," in *2025 Global Conference in Emerging Technology (GINOTECH)*, IEEE, May 2025, pp. 1–6. doi: 10.1109/GINOTECH63460.2025.11076855.

[45] Y. L. M. Latha, S. Choudhuri, A. R. Jannepally, Y. D. Bhise, T.

Swathi, and S. Rao Katta, "Retracted: Optimizing Cloud Resource Management Through AI Enhancing Efficiency and Scalability in Modern Computing Environments," in *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*, IEEE, Sep. 2024, pp. 1044–1049. doi: 10.1109/GlobalAISummit62156.2024.10947796.

[46] J. Barach and J. H. Muzamal, "Enhancing Resource Management and Energy Consumption Forecasting in Multi-Cloud Environments with AI-driven Approaches," in *2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, IEEE, Oct. 2024, pp. 1–8. doi: 10.1109/DESSERT65323.2024.11122186.

[47] S. Nagasundaram, B. N. Bobinath, A. Shedthi, K. Rajalakshmi, T. D. Humnekar, and Millerjothi, "Analysis of the Requirement and Artificial Intelligence-based Resource Management System in Cloud," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2023, pp. 2516–2525. doi: 10.1109/ICACCS57279.2023.10112940.