SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I5P106

Volume 12 Issue 5, 68-79, May 2025

Review Article

A Comprehensive Survey on Energy and Performance-Aware Scheduling Approaches in Cloud Computing

Anil Kumar D B[1,3], Raghu N[2]

[1,2]*Faculty of Engineering and Technology, JAIN (Deemed to be University), Karnataka, India.*
[3]*School of Electrical and Electronics Engineering, REVA University, Karnataka, India.*

[1]*Corresponding Author : anilci.12@gmail.com*

Received: 04 March 2025 Revised: 06 April 2025 Accepted: 07 May 2025 Published: 27 May 2025

Abstract *- As cloud computing is growing rapidly and resource management is becoming more important, energy and performance-aware task scheduling in cloud computing is an important study topic. To optimize energy usage and boost overall system performance in cloud computing, this study gives a comprehensive overview of the available research on task scheduling strategies. We show the advantages and limitations of a wide variety of scheduling algorithms, methods, and approaches. This review examines how academics and industry professionals have dealt with cloud computing's unique mix of difficulties, including fluctuating workloads, diverse resource requirements, and unpredictable performance. Important gains in both energy economy and system performance have resulted from adopting machine learning techniques that can adapt to various scheduling conditions. Our research highlights the need for data-driven models and Artificial Intelligence (AI-assisted) scheduling decisions to maximize resource utilization and satisfy the wide-ranging performance needs of cloud applications. We also highlight areas where additional study is needed, such as in the areas of large-scale data processing, security, and dynamic scheduling systems that can accommodate workload changes. This survey is useful for researchers, professionals, and policymakers in the domain of Cloud Computing (CC) since it compiles the results of a wide range of studies.*

Keywords *- Energy and performance aware scheduling, Cloud computing, Quality of service, Dynamic workload, Cache aware scheduling.*

1. Introduction

Eric Schmidt, former CEO of Google, was mainly responsible for bringing cloud computing to the public in 2006. Since then, it has been a key technology driving the expansion of data-intensive applications in domains such as scientific research, industry, and consumer services. Cloud architecture has caused a paradigm shift by providing scalable access to massive computing resources online, enabling applications that require large processing power and storage space [1]. Cloud services allow users to acquire computer resources on demand, paying based on actual consumption as specified in their Service Level Agreements (SLAs), providing both flexibility and cost savings. The paradigm comprises layers such as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS), which cater to a wide range of customer demands while requiring providers to maintain high levels of service quality, availability, and adaptive resource management [3]. The different cloud service models are illustrated in Figure 1. A CSP will route user requests to any of its available resources and then provide the outcome to the requesting user. A request's resource allocation is determined on the fly. Systematic resource management ensures continuous service and optimal system performance. Cloud computing has been successful because of its scalability and the fact that it can provide computing and storage resources on demand. With cloud computing, one may access computer resources on specific requirements by paying for them on a usage basis. Users in a Cloud model are charged according to their resource consumption and the Quality of Service (QoS) requirements, as defined in the Service Level Agreement (SLA) among the Cloud provider and the user [3]. There are essentially two participants in the cloud setup: the resource supplier and the consumer. Users seek to perform their operations in the form of workflows or independent tasks in a timely way with a minimum runtime cost, while providers strive to maximize their advantages from cloud infrastructure [4, 5].

Cloud data centres are seeing exponential increases in resource use as demand for cloud services rises. Along with this expansion, energy usage has surged, which raises environmental and financial issues. Data centres presently consume around 200 TWh yearly, a figure that could reach 3,000 TWh by 2030 if present trends continue, therefore casting questions on the sustainability of cloud infrastructure [6]. Consequently, in cloud computing research, energy-

efficient job scheduling and resource management techniques have been essential. Optimizing both energy consumption and performance depends critically on task scheduling, which determines how and when resources are distributed for performing different activities. All of these are vital for the sustainability of cloud services, as they cover attempts to maximize system efficiency, lower idle resource time, and cut wasteful power use.

Fig. 1 Different cloud service models

To tackle these challenges, researchers have developed various task-scheduling algorithms aimed at boosting energy efficiency while sustaining or even improving system performance. These approaches range from traditional heuristic methods to newer machine learning techniques that can dynamically adjust to changing workloads [7-9]. Yet, a notable gap in current research is the scalability and flexibility of these models across diverse workloads and operational settings. Scheduling scientific workflows, especially in fields like bioinformatics, climate modeling, and high-energy physics, poses distinct challenges due to their complex resource requirements, task interdependencies, and often time-sensitive constraints [7]. Traditional scheduling methods often struggle with these specialized workflows, as they typically do not meet specific performance needs, energy-saving objectives, or the dynamic nature of such tasks.

In this area, a notable research gap becomes evident. Despite the development of numerous task scheduling strategies, there is still a need for comprehensive solutions that effectively balance energy efficiency with performance goals, particularly in large-scale scientific workflows. Current methods often fall short of adapting to fluctuating workload demands and struggle to achieve low energy use and high computational efficiency simultaneously. Additionally, while AI and data-driven scheduling models offer potential for real-time, adaptive scheduling, their full capabilities for enhancing energy efficiency and optimizing resource use in cloud environments remain largely unexplored.

This paper seeks to bridge this gap by offering a detailed review of recent developments in energy-efficient, performance-oriented scheduling methods within cloud computing. Through systematic analysis aimed to explore how AI-assisted scheduling models and adaptive, data-driven strategies can enhance resource utilization, meet quality of service standards, and improve overall energy efficiency. By assessing the strengths and limitations of these techniques, this survey aims to highlight the most effective approaches for balancing energy and performance needs in cloud data centres, particularly for scientific workflows with distinct operational demands.

2. Review Methodology

To maintain scientific rigor and thoroughly cover energy and performance-aware scheduling in cloud computing, a systematic review methodology was used, adhering to recognized guidelines for conducting structured literature surveys. This section outlines the approach taken, including the databases consulted, keywords used, and criteria applied for selecting studies to include.

To gather a comprehensive range of peer-reviewed studies and relevant sources, we searched multiple high-impact databases, including “IEEE Xplore”, “ScienceDirect”, “SpringerLink”, “Wiley Online Library”, and “Google Scholar” (for additional coverage). These databases were selected for their extensive collections of computer science and engineering journals, with a strong focus on cloud computing and related technologies.

Used a combination of keywords and phrases to conduct a thorough search for relevant studies. The main search terms included:

- Cloud Computing Task Scheduling.
- Energy-aware Scheduling Algorithms.
- Performance Optimization in Cloud Computing.
- Heterogeneous Computing Environments in Cloud.
- Real-time Scheduling in the Cloud.
- Dynamic Resource Allocation Cloud.
- AI-driven Scheduling in Cloud Computing.
- Sustainable Cloud Data Centres.

Boolean operators (AND, OR) were applied to combine these keywords, allowing us to adjust the breadth or specificity of search results as needed.

Studies were chosen based on criteria designed to ensure relevance and focus, including:

- Publications from the last ten years to reflect current technologies and methodologies.

- Peer-reviewed articles, conference papers, and reputable surveys centred on energy-efficient or performance-focused scheduling algorithms in cloud computing.
- Studies covering both homogeneous and heterogeneous cloud environments and those discussing real-time scheduling or fault tolerance.

The selected studies were categorized based on their focus areas, including energy-aware scheduling, performance-driven scheduling, and adaptive scheduling within heterogeneous environments. An in-depth analysis was performed to uncover recurring trends, technological gaps, and opportunities for future research. A comparative analysis table was also developed to showcase various scheduling approaches' methodologies, limitations, and performance outcomes, offering a structured summary of the findings.

3. Energy Aware Scheduling Algorithms

The authors offer a scheduling technique for a workflow scheduling issue under time constraints and geographically dispersed data that minimises power consumption [8]. In addition, the authors discuss methods for arranging workflow applications, setting priorities, and prioritising work. Experimental findings show that the suggested method is superior to competing algorithms. The paper stresses the need to minimise energy use throughout the workflow scheduling process due to the rising power usage of cloud data centres. For the issue of cloud data centre energy usage, [9] suggested a novel model for processing applications effectively and a scheduling method based on Particle Swarm Optimisation (PSO) that minimises cost and execution time by taking delay into account. In contrast, a linear programming-based resource allocation approach was described in [10] for effectively computing high-quality solutions that minimise energy and make-span concurrently. The proposed approach attempts to find a suitable compromise amongst resource conservation and productivity. Workflow execution on a DVFS-enabled cloud is presented, together with the Smart Energy and Reliability Aware Scheduling algorithm (SERAS) [11]. This approach's goal is to minimise power usage without compromising on dependability or completion time. This research emphasises the significance of energy usage in a green cloud setting and the use of the DVFS approach in this regard. The authors suggested a two-stage technique that uses less power and finishes faster. Algorithmically, the graph of tasks with precedence constraints is duplicated and clustered so that they may be scheduled on data centre processors using DVFS [12].

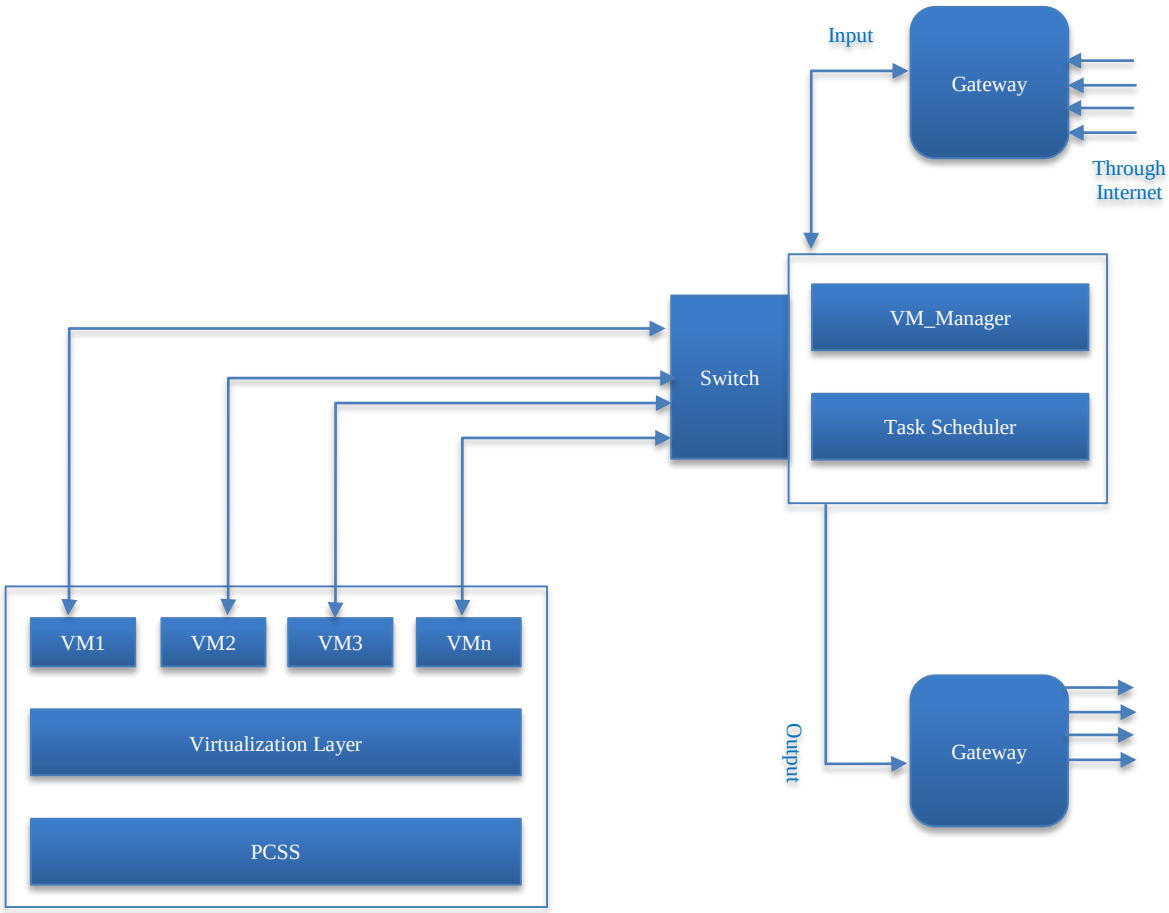

Fig. 2 Distributed resource management scheme [18]

The work presents a "combinatorial auction-based" solution for the resource allocation issue that takes energy efficiency into account. The suggested approach lets cloud users place bids on the virtual resources they need by utilising a bidding language that lets them specify the complementarities and substitutabilities among the resources they are competing on [13].

The model employs an optimisation problem to identify the most profitable set of winning bids and the related allocation of virtual resources to users, considering the deployment of virtual resources relative to cloud physical resources. This research used CPU power utilisation models to allocate virtual machines to optimise server utilisation and minimise energy consumption. Dynamic virtual machine consolidation reduces infrastructure power usage while preserving SLAs [14-16]. Work in [17] proposes a VM placement architecture that minimises power usage, maintains service quality, and adheres to performance and security targets.

This paper discusses the difficulties of optimising cloud computing for energy efficiency, dependability, and affordability in the context of scientific processes. The authors provide a system to facilitate these aims, together with the optimisation of costs, fault tolerance, and operational efficiency [18]. They stress the need for dispersed resource management and service quality improvement to meet these difficulties. The energy-aware method is depicted in Figure 2, which entails characteristics such as dynamic LAN, VM, and data storage switching. [19] proposes dividing the scope of the rational process in accordance with available resources to reduce workflow and resource costs. Designing and building an accurate model is challenging because of the need for failure data to achieve a certain goal and because companies are reluctant to disclose failure data because of privacy concerns.

In [20], an NSGA-III-based multi-objective workflow scheduling optimisation framework was presented to consider the importance of data transportation in determining the efficiency and sustainability of network equipment in cloud data centres. Using extreme solutions in population initialization, the proposed technique improves solution quality. Popular scientific procedures are chosen to serve as testbeds. The goals pursued by each of these works also allow us to differentiate between them. The most researched primary goals are decreasing the make-span/execution time, energy utilization, and the execution cost of the scientific workflows. Multiple permutations of these goals are of particular interest. As can be shown in Figure 3, the primary goals of the relevant literature are make-span, cost, and energy. Nearly all of the connected works employ the cloud simulation tool, as seen in Figure 4 consumption, the efficiency of the programme consumption, and the efficiency of the programme. It may be divided into headed subsections if several methods are described. The article proposes a simple task.

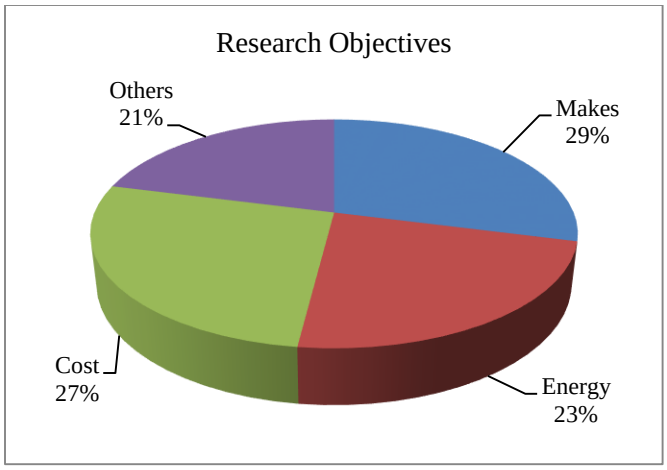

Fig. 3 Research objectives in the past years

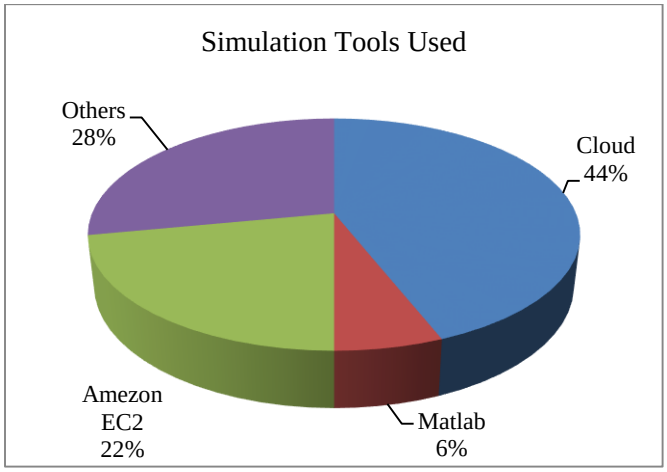

Fig. 4 Simulation tools used

4. Performance-Aware Scheduling Techniques

Scientific computing spans many different fields of study, such as high-energy physics, molecular dynamics simulations, and climate modelling. Large amounts of memory, storage capacity, and the ability to coordinate several tasks are general requirements of such applications. These specialised scientific tasks may not best fit traditional scheduling methods and heuristics built for general-purpose workloads. Scheduling scientific workloads on the cloud while keeping performance goals in mind is an acute issue that must be solved. Possible metrics to track include task completion time, resource use, and energy computing environments [21]. The resource selection process in a cloud/edge arrangement may be completed with minimal extra effort using the suggested framework based on Reliable Server Pooling (RSerPool), shown in Figure 5. The article explains how offloading work to the cloud and edge might boost application performance, considering the rising popularity of mobile devices. The availability of resources and the nature of the workload are the primary considerations when deciding whether to offload work from mobile nodes to edge or even multi-cloud resources.

Table 1. Comparative analysis

Paper	Objective	Method	Tools	Constraints	Remarks
[8]	Energy Aware.	Adaptive local search.	Cloud-sim.	Deadline.	Geo-distributed data.
[9]	Energy Aware.	Particle-Swarm Optimization.	Cloud-sim.	Deadline.	Resource-allocation model.
[10]	Energy Aware.	Linear Programming Technique.	Amazon EC2.	Budget.	Resource allocation algorithm.
[11]	Energy Aware.	DVFS technique.	Cloud-sim.	-	Workflow Execution.
[12]	Energy Aware.	DVFS technique.	Cloud-sim.	Deadline.	Parallel Task-based applications.
[13]	Energy Aware.	Combinatorial auction-based model.	Amazon EC2.	Deadline.	Green Cloud.
[14]	Energy Aware.	Modified Best Fit Decreasing.	Cloud-sim.	-	Resource-allocation model.
[15]	Energy Performance Trade-off.	Heuristic-Based Approach, fast best-fit Decreasing.	-	Deadline.	VM Consolidation.
[16]	Energy Aware.	Multiple-linear regression model.	Amazon EC2.	-	Dynamic provisioning of resources.
[18]	Energy & Cost.	Adaptive-Cloud Resource Re-Configurability.	Matlab.	-	Workflow scheduling.
[20]	Energy & Cost.	Multilevel-Dependent Node Clustering.	Amazon EC2.	Deadline.	Multi-objective. scheduling.

To dynamically join or leave a pool, Pool Elements (Pes) utilise the "Aggregate Server Access Protocol" (ASAP) to communicate with a pool register in their area of operation [22]. The paper discusses about an energy-efficient offloading strategy for IoT applications in a "Fog-Cloud environment". The strategy uses a Firefly algorithm to obtain an optimal computing device depending on energy utilization, execution time, and QoS parameters. The study argues that cloud and Fog computing must cooperate to create a long-term IoT architecture. Given the wide variety of IoT applications and the need for powerful computing units, task offloading presents a difficult challenge in a Fog-Cloud setting [23].

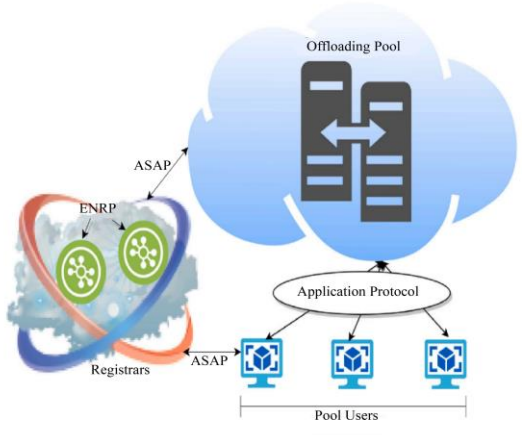

Fig. 5 Pool architecture

Edge computing is discussed as a solution for the instantaneous processing of tasks locally and the difficulties of managing the large amounts of data created by IoT devices. Authors propose a method called Priority aware Task arranging (PaTS) to unload data from edge and cloud servers by arranging jobs based on their priority [24]. The efficiency of the presented technique, which is conceptualised as a multi-objective function, is evaluated using the bio-inspired NSGA-2.

Using Density-based spatial clustering, this research proposes a task scheduling technique for cloud computing, optimising resource utilisation and minimising duplication of effort [25]. The method's goal is to speed up the beginning, middle, and end of user-initiated actions. The suggested model improves upon prior art Ant Colony Optimization (ACO) and PSO algorithms by a margin of 13% in terms of execution time and 49% in terms of both average start time and average finish time.

The algorithm's goal is optimal throughput and efficiency. A multi-objective scheduling approach was presented for scheduling tasks on cloud infrastructure by the authors of [26]. The model was conceptualised from observing how squirrels navigate their environment in quest of food.

Performance criteria such as energy consumption, cost, and utilisation are used to evaluate the proposed model against a Bat inspired model, PSO, and a hybrid genetic algorithm. As

the number of jobs rises, the suggested model outperforms state-of-the-art approaches in regard to energy efficiency and cloud performance.

In this paper, the authors provide a hybrid paradigm for scheduling cloud-based tasks [27]. The model is built on both neural networks and meta-heuristics. Using performance measurements, the model determines the best way to distribute work among cloud resources. When compared to current meta-heuristic methods, the suggested DE-ANN methodology performs better.

Two setup scenarios are used to assess performance: one with five virtual machines and another with ten virtual machines and a range of 1000 to 4500 tasks. Mahmoud et al. have stressed the potential of ML techniques for multi-objective task scheduling in the cloud. Applications that need low system overhead and power consumption, such as those that use cloud computing, may benefit from the proposed hybrid algorithm's potential to construct efficient work schedules that maximise many objectives concurrently [28].

The paper discusses a technique for scheduling tasks in edge computing settings that takes cache storage into consideration [29]. To maximise caching value while minimising cache replacement penalty, the suggested technique first derives an integrated utility function before moving on to a cache locality-based task scheduling approach modelled as a weighted bipartite network. However, the proposed algorithms beat the best-existing baseline methods while maintaining polynomial time complexity in all these areas: cache hit ratio, data locality, data transmission time, task response time, and energy consumption costs. By running some applications on edge servers, edge computing can lower both latency and energy costs.

Within the context of Mobile-Edge Computing (MEC) networks, this article discusses a suggested collaborative joint caching and processing technique for on-demand video streaming. By strategically caching different bitrate versions of videos and taking into account transcoding relationships among different versions, the plan aims to enhance the current Adaptive Bitrate (ABR) streaming technology. Results from many simulations show that the suggested method significantly outperforms the state-of-the-art approaches in cache hit ratio, backhaul traffic, and first-access time [30].

The paper delves into how cloud-based video distribution systems might better optimise data-intensive applications. High data availability at a cheap storage cost is the goal of the new cloud file system CAROM proposed by the authors. The suggested method utilises a tripartite graph and a k-list algorithm as an approximation technique to define the interconnections between tasks, computation nodes, and data nodes. The suggested approach outperforms the standard two-layer algorithm in simulations [31]. The paper discusses an algorithm for efficient resource management in cloud data centres that takes memory into consideration. This solution provides a memory priority cloud task mapping rule that considers the cloud job's CPU and memory characteristics to decrease C cloud customers' prices and data centre energy usage.

Green cloud computing and energy efficiency in the cloud are discussed, emphasising the former [32]. Using the combined power of many data centres, the study details how to carry out data-intensive scientific activities in the cloud. The authors provide a solution for cache-aware scheduling of scientific workflows in a multisite cloud by utilising a distributed and parallel architecture and novel approaches for adaptive caching, cache site selection, and dynamic workflow scheduling. The experimental assessment demonstrates that the suggested technique may significantly improve performance, decreasing total time by up to 42% while using 60% of the original input data [33].

4.1. Challenges Faced in Task Scheduling

Cloud data centres are naturally diverse, featuring a range of configurations in hardware, operating systems, and virtualized resources. This variation makes task scheduling more complex, as algorithms must adjust to resource capabilities, energy efficiency, and workload management differences. This section will explore how diverse hardware affects the balance between energy consumption and performance, particularly in multi-cloud or hybrid environments.

Many applications depend on real-time processing, requiring tasks to be executed instantly with minimal delay. Balancing energy efficiency under such tight timing demands is challenging, as real-time applications need swift resource allocation, which can increase energy use to prevent delays. I will explain how task scheduling solutions can meet real-time requirements while still supporting energy-saving goals.

Cloud environments experience continuous fluctuations in workload demand, requiring adaptive scheduling strategies to respond to real-time changing resource needs. This variability often results in either underutilized or overused resources, making it challenging to sustain optimal energy efficiency. Section 3 examines how these dynamic conditions complicate efficient resource allocation and proposes adaptive approaches to help balance energy use with performance demands.

Maintaining reliable service while reducing energy consumption is further complicated by the need for fault tolerance. Many fault tolerance mechanisms require extra resources, which can raise energy overhead. This section will explore common methods for achieving fault tolerance in cloud computing and examine their impact on energy-efficient scheduling strategies.

Table 2. Comparative analysis

Paper	Objective	Method	Tools	Constraints	Remarks
[21]	Task Scheduling.	ReliableServer Pooling.	RSPSIM simulation model.	Resource utilization.	The RSerPool framework may function satisfactorily in a cloud environment.
[23]	Task Offloading.	Firefly algorithm.	-	Execution Time.	Multi-optimization strategy.
[25]	Task Scheduling.	Density-based spatial clustering.	Cloud-sim.	Execution Time.	The performance of the method can be enhanced by combining it with a machine learning algorithm.
[26]	Task Scheduling.	Chaotic squirrel search algorithm.	Cloud-sim.	QoS.	Multi-optimization strategy.
[27]	Task Scheduling.	Differential Evolution.	-	Resource utilization.	Task distribution among many virtual machines.
[29]	Memory Aware.	Weighted bipartite graph.	-	Cache Allocation.	Data is cached in pieces at the most efficient edge servers.
[32]	Memory Aware.	Male algorithm.	Cloud-sim.	Resource Utilization.	The technique uses a mechanism for mapping memory priorities onto cloud tasks to speed up routine operations.
[33]	Memory Aware.	Distributed and parallel architecture.	Open-Alea workflow system.	Cache Allocation.	Workflow scheduling and interim data caching.

5. Combined Energy and Performance-Aware Scheduling Techniques

The article discusses a bi-objective method for optimising cloud data centre performance and energy consumption. The technique is tested with an actual cloud dataset, and its performance is optimised through evolutionary multi-objective optimisation based on system performance counters. The essay stresses the need to maintain QoS in smart city services and applications [34]. The paper discusses the Harmony-Inspired Genetic Algorithm (HIGA), a novel hybrid metaheuristic system for scheduling tasks in cloud data centres in a way that minimises energy use. The technique utilises both the genetic algorithm's exploration capabilities and the harmony search's exploitation capabilities to rapidly converge on globally optimum solutions. The primary objectives are reducing the makespan, the amount of processing energy used, the resources required, and the execution overhead. The simulation results suggest that HIGA can reduce energy consumption by up to 33%, boost application performance by up to 47%, and reduce execution overhead by up to 39% [35]. The authors discuss a trust-aware, multi-objective, task-scheduling algorithm for the cloud that makes use of whale optimisation. The algorithm considers trust criteria, including availability, success rate, and turnaround efficiency, while it works to reduce makespan and energy usage. The simulation findings demonstrate dramatic gains over prior metaheuristic methods.

Maintaining quality of service and trust between cloud users and service providers is a key focus of this essay, highlighting the need for efficient scheduling in cloud computing [36, 37]. To better allocate virtual machines to physical computers in cloud data centres and cut down on energy usage and resource utilisation, the paper describes a hybrid optimisation technique named FHCS. The technique is tested with Cloud-Sim simulation and is a hybrid of the Fruit Fly Optimisation and Cuckoo Search algorithms [38]. The article offers a cloud-based task-scheduling algorithm that uses best-worst and TOPSIS multi-criteria decision-making to reduce power consumption without sacrificing users' quality-of-service requirements. The algorithm has five steps: user task submission, establishing assessment criteria, ranking tasks, giving important weights, and a dispatcher that maximises energy efficiency. Several benchmarks are used to

compare the proposed methodology to others; the results demonstrate that the suggested method efficiently decreases makespan and energy consumption while increasing VM utilisation. Table 3 compares various bi-objective approaches.

6. Case Study Examples

6.1. Energy-aware Resource Provisioning using DVFS Technique

As the importance of energy efficiency in data centre management has grown, new methods have emerged to maximise efficiency while cutting power usage. DVFS is one such method that has received a lot of interest. DVFS is a technique for saving energy that modifies a CPU's voltage and frequency in response to its current task. A processor's ability to dynamically adjust these settings allows it to optimise its performance and energy consumption for the running task. Using the DVFS method, the authors propose a method known as "multi-agent deep Q-network" with "coral reefs optimisation" (MDQ-CR), which is a blend of Coral Reefs Optimisation algorithm (CRO) and multi-agent deep Q-network. The learning agents are guided to the global optimum solution with the help of the Markov game model [40]. Recent bio-inspired optimisation algorithms like the CRO algorithm are useful in various contexts, including cloud-based resource management [41]. The algorithm has a good convergence rate and speed, and it was inspired by the coral reef ecosystem in the ocean. The CRO algorithm's flowchart is seen in Figure 6. Due to its learning-based model and parallelization, the disclosed method meets these objectives with outstanding precision. Overhead time due to selecting an insufficient number of parallel agents is a drawback of the suggested approach. Since the Markov game model converges parallel agents, allocating resources to more than one threshold takes more time.

6.2. Scientific Workload Scheduling Using Resource Allocation Technique

The authors provide a method for effectively allocating resources while planning scientific workload in the cloud [42]. To satisfy the timeliness requirements of workload tasks while making optimal use of cache resources, the authors suggest an Efficient Resource Utilisation (ERU) model. Execution time, power usage, and energy usage are all reduced when using the ERU model to run scientific workflows on a heterogeneous cloud environment, compared to the conventional way of resource management. The essay stresses the significance of effective resource scheduling and the difficulties of controlling power consumption in multicore devices. The Last Level Cache (LLC) failure rate is kept to a minimum in a shared caching environment because of the ERU's clever architecture. To alleviate LLC failures and satisfy the cache restriction, VM migration is performed in this case. The "short ribonucleic acid" (sRNA) Identification Procedure Utilising High-Throughput Technology (SIPHT) in a scientific process is used to validate the high efficiency and low energy consumption of the presented cache-aware resource energy use. A cache-aware efficient resource utilisation scientific workflow scheduling approach is developed to guarantee low energy consumption, good model performance, and optimal resource scheduling while employing heterogeneous multi-core architectures. This method improves the model's speed and efficiency. A summary of scheduling algorithms in cloud computing is discussed in Table 4.

Table 3. Comparative Analysis of the bi-objective Approach

Paper	Methodology	Tools	Constraints	Remarks
[34]	Evolutionary algorithm, Modified Worst Fit Decreasing.	Cloud-sim.	Energy consumption and mean execution time.	Performance & Energy Optimization.
[35]	Harmony-inspired genetic algorithm.	Cloud-sim.	Energy consumption and mean execution time.	Maximising performance and efficiency with minimal resources.
[36]	Whale optimization.	Cloud-sim.	Energy-consumption, Turnaround efficiency.	Scheduler that considers workloads, virtual machines, and the allocation of available virtual resources.
[39]	Technique for Order Preference by Similarity to Ideal Solution.	Cloud-sim.	Service-level agreement.	TOPSIS takes the weighted criteria as inputs to rate and rank the quality of each feasible choice.

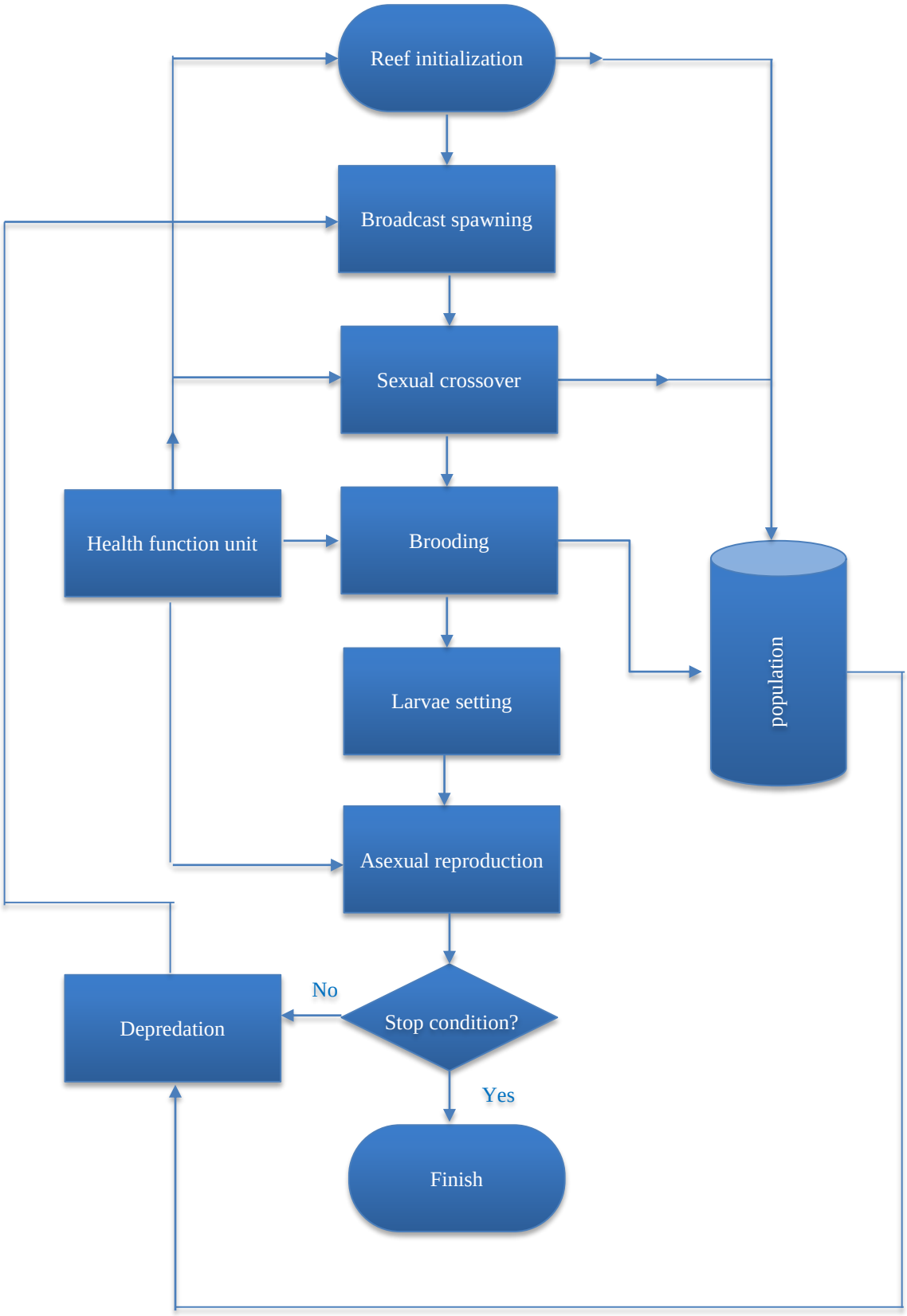

Fig. 6 CRO algorithm flowchart [41]

Table 4. Comparative analysis of bi-objective approach

Reference Paper	Methodology	Evaluation Metrics	Limitations
[1]	Data-aware scheduler.	Execution Time.	Other QoS factors, such as makes pan and energy usage, are ignored by the proposed approach.
[6]	DVFS technique.	Finish Time.	It does not consider efficiency in terms of cost.
[9]	PSO Algorithm.	Execution Time, Finish Time.	Didn't reconsider how occupied a server already was or user priority. Cost and energy of execution are ignored
[12]	DVFS technique.	Energy-Consumption, makes pan.	Things like the cost of execution, the percentage of rejected tasks, the throughput, and so on. The suggested method does not take quality-of-service factors into account.
[20]	Multilevel Dependent Node Clustering.	Latency.	Energy consumption, renewable energy sources, and carbon dioxide emissions are not accounted for in these multi-objective problems.
[24]	Priority aware Task Scheduling.	Latency, Execution Time.	Superior performance was shown on a small sample of simulated data.
[28]	Task-Scheduling-Decision Tree.	Make span, Load Balance.	Some crucial metrics, such energy efficiency, fault tolerance, and scalability, are ignored.
[33]	Distributed caching.	Execution Time.	Distributed scheduling methods are concerned with improving workflow execution, none of them take intermediate data caching and reuse into account.

7. Conclusion

The objective of the study was to undertake a systematic literature review on the critical issue of scheduling tasks in cloud computing infrastructures in a way that maximises efficiency while also optimising performance. The need for effective task scheduling strategies to optimise resource utilisation, minimise energy consumption, and boost overall system performance has increased as cloud services have become more widely used and data demands have grown exponentially. Through a systematic analysis of different task scheduling algorithms and methods, this survey shows the different ways researchers and practitioners have tried to deal with the problems caused by different kinds of resources, changing workloads and different needs for performance in cloud environments. The surveyed literature encompasses a wide array of techniques, ranging from traditional heuristic-based algorithms to sophisticated machine-learning-based approaches. The primary findings from this review reveal that, while classical algorithms are still effective in some contexts, machine learning approaches have shown a lot of promise in dealing with complicated scheduling challenges. Energy economy and system performance improvements have resulted from smarter, more adaptable scheduling decisions made possible by using artificial intelligence and data-driven models. In addition, the article pointed out a number of open research difficulties in this area, including dealing with massive amounts of data, being mindful of privacy and security issues, and developing scheduling techniques that can accommodate fluctuating workloads. Resolving these issues is essential to the continued success of cloud computing because it will spur innovation in energy and performance-aware task scheduling.

References

[1] Salvatore Giampà et al., "A Data-Aware Scheduling Strategy for Executing Large-Scale Distributed Workflows," *IEEE Access*, vol. 9, pp. 47354-47364, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] M. Menaka, and K.S. Sendhil Kumar, "Workflow Scheduling in Cloud Environment – Challenges, Tools, Limitations & Methodologies: A Review," *Measurement: Sensors*, vol. 24, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Pham Phuoc Hung, and Eui-Nam Huh, "An Adaptive Procedure for Task Scheduling Optimization in Mobile Cloud Computing," *Mathematical Problems in Engineering*, vol. 2015, no. 1, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Guoqi Xie et al., "A Survey of Low-Energy Parallel Scheduling Algorithms," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 1, pp. 27-46, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Vrunda J. Patel, and Hitesh A. Bheda, "Reducing Energy Consumption with DVFS for Real-Time Services in Cloud Computing," *IOSR Journal of Computer Engineering*, vol. 16, no. 3, pp. 53-57, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zhongjin Li et al., "Cost and Energy Aware Scheduling Algorithm for Scientific Workflows with Deadline Constraint in Clouds," *IEEE Transactions on Services Computing*, vol. 11, no. 4, pp. 713-726, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] Michel Krämer, Hendrik M. Würz, and Christian Altenhofen, "Executing Cyclic Scientific Workflows in the Cloud," *Journal of Cloud Computing*, vol. 10, pp. 1-26, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Xiaoping Li et al., "Energy-Aware Cloud Workflow Applications Scheduling with Geo-Distributed Data," *IEEE Transactions on Services Computing*, vol. 15, no. 2, pp. 891-903, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mohit Kumar, and S.C. Sharma, "PSO-COGENT: Cost and Energy Efficient Scheduling in Cloud Environment with Deadline Constraint," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 147-164, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Kyle M. Tarplee et al., "Energy and Makespan Tradeoffs in Heterogeneous Computing Systems Using Efficient Linear Programming Techniques," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 6, pp. 1633-1646, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[11] Hadeer A. Hassan, Sameh A. Salem, and Elsayed M. Saad, "A Smart Energy and Reliability Aware Scheduling Algorithm for Workflow Execution in DVFS-Enabled Cloud Environment," *Future Generation Computer Systems*, vol. 112, pp. 431-448, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] B. Barzegar, H. Motameni, and A. Movaghar, "EATSDCD: A Green Energy-Aware Scheduling Algorithm for Parallel Task-Based Application Using Clustering, Duplication and DVFS Technique in Cloud Datacenters," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 36, no. 6, pp. 5135-5152, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mustafa Gamsiz, and Alí Haydar Özer, "An Energy-Aware Combinatorial Virtual Machine Allocation and Placement Model for Green Cloud Computing," *IEEE Access*, vol. 9, pp. 18625-18648, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya, "Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755-768, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[15] Monir Abdullah et al., "A Heuristic-Based Approach for Dynamic VMs Consolidation in Cloud Data Centers," *Arabian Journal for Science and Engineering*, vol. 42, pp. 3535-3549, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[16] Célia Ghedini Ralha et al., "Multiagent System for Dynamic Resource Provisioning in Cloud Computing Platforms," *Future Generation Computer Systems*, vol. 94, pp. 80-96, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mohammad Masdari, and Mehran Zangakani, "Green Cloud Computing Using Proactive Virtual Machine Placement: Challenges and Issues," *Journal of Grid Computing*, vol. 18, pp. 727-759, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sudha Danthuluri, and Sanjay Chitnis, "Energy and Cost Optimization Mechanism for Workflow Scheduling in the Cloud," *Materials Today: Proceedings*, vol. 80, no. 3, pp. 3069-3074, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Reihaneh Khorsand et al., "Taxonomy of Workflow Partitioning Problems and Methods in Distributed Environments," *Journal of Systems and Software*, vol. 132, pp. 253-271, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] Peerasak Wangsom, Kittichai Lavangnananda, and Pascal Bouvry, "Multi-Objective Scientific-Workflow Scheduling with Data Movement Awareness in Cloud," *IEEE Access*, vol. 7, pp. 177063-177081, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Thomas Dreibholz, and Somnath Mazumdar, "Towards a Lightweight Task Scheduling Framework for Cloud and Edge Platform," *Internet of Things*, vol. 21, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Beneyaz Ara Begum, and Satyanarayana V. Nandury, "Data Aggregation Protocols for WSN and IoT Applications – A Comprehensive Survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 651-681, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mainak Adhikari, and Hemant Gianey, "Energy Efficient Offloading Strategy in Fog-Cloud Environment for IoT Applications," *Internet of Things*, vol. 6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[24] Malvinder Singh Bali et al., "An Effective Technique to Schedule Priority Aware Tasks to Offload Data on Edge and Cloud Servers," *Measurement: Sensors*, vol. 26, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] S.M.F.D. Syed Mustapha, and Punit Gupta, "DBSCAN Inspired Task Scheduling Algorithm for Cloud Infrastructure," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 32-39, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] M.S. Sanaj, and P.M. Joe Prathap, "Nature Inspired Chaotic Squirrel Search Algorithm (CSSA) for Multi Objective Task Scheduling in an IAAS Cloud Computing Atmosphere," *Engineering Science and Technology, An International Journal*, vol. 23, no. 4, pp. 891-902, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Punit Gupta et al., "Neural Network Inspired Differential Evolution Based Task Scheduling for Cloud Infrastructure," *Alexandria Engineering Journal*, vol. 73, pp. 217-230, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Hadeer Mahmoud et al., "Multiobjective Task Scheduling in Cloud Environment Using Decision Tree Algorithm," *IEEE Access*, vol. 10, pp. 36140-36151, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Chunlin Li et al., "Collaborative Cache Allocation and Task Scheduling for Data-Intensive Applications in Edge Computing Environment," *Future Generation Computer Systems*, vol. 95, pp. 249-264, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[30] Tuyen X. Tran et al., "Collaborative Multi-Bitrate Video Caching and Processing in Mobile-Edge Computing Networks," *2017 13th Annual Conference on Wireless On-demand Network Systems and Services*, Jackson, WY, USA, pp. 165-172, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] Xili Dai, Xiaomin Wang, and Nianbo Liu, "Optimal Scheduling of Data-Intensive Applications in Cloud-Based Video Distribution Services," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 1, pp. 73-83, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] Bin Liang et al., "Memory-Aware Resource Management Algorithm for Low-Energy Cloud Data Centers," *Future Generation Computer Systems*, vol. 113, pp. 329-342, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[33] Gaëtan Heidsieck et al., "Cache-Aware Scheduling of Scientific Workflows in a Multisite Cloud," *Future Generation Computer Systems*, vol. 122, pp. 172-186, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[34] Huned Materwala, and Leila Ismail, "Performance and Energy-Aware Bi-Objective Tasks Scheduling for Cloud Data Centers," *Procedia Computer Science*, vol. 197, pp. 238-246, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[35] Mohan Sharma, and Ritu Garg, "HIGA: Harmony-Inspired Genetic Algorithm for Rack-Aware Energy-Efficient Task Scheduling in Cloud Data Centers," *Engineering Science and Technology, an International Journal*, vol. 23, no. 1, pp. 211-224, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[36] Sudheer Mangalampalli, Ganesh Reddy Karri, and Utku Kose, "Multi Objective Trust Aware Task Scheduling Algorithm in Cloud Computing Using Whale Optimization," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 791-809, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] A.S. Ajeena Beegom, and M.S. Rajasree, "Integer-PSO: A Discrete PSO Algorithm for Task Scheduling in Cloud Computing Systems," *Evolutionary Intelligence*, vol. 12, pp. 227-239, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[38] Banavath Balaji Naik, Dhananjay Singh, and Arun B. Samaddar, "FHCS: Hybridised Optimisation for Virtual Machine Migration and Task Scheduling in Cloud Data Center," *IET Communications*, vol. 14, no. 12, pp. 1942-1948, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[39] Reihaneh Khorsand, and Mohammadreza Ramezanpour, "An Energy-Efficient Task-Scheduling Algorithm Based on a Multi-Criteria Decision-Making Method in Cloud Computing," *International Journal of Communication Systems*, vol. 33, no. 9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Ali Asghari, and Mohammad Karim Sohrabi, "Combined Use of Coral Reefs Optimization and Multi-Agent Deep Q-Network for Energy-Aware Resource Provisioning in Cloud Data Centers Using DVFS Technique," *Cluster Computing*, vol. 25, pp. 119-140, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[41] A. Asghari, and M.K. Sohrabi, "Combined Use of Coral Reefs Optimization and Reinforcement Learning for Improving Resource Utilization and Load Balancing in Cloud Environments," *Computing*, vol. 103, pp. 1545-1567, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[42] Nagendra Prasad Sodinapalli et al., "An Efficient Resource Utilization Technique for Scheduling Scientific Workload in Cloud Computing Environment," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 367-378, 2022. [CrossRef] [Google Scholar] [Publisher Link]