International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

A REVIEW OF ENERGY-EFFICIENT TASK SCHEDULING TECHNIQUES IN CLOUD COMPUTING ENVIRONMENTS

Raj Luxmi Yadav[1], Mr. Rajesh Sharma[2]

[1]*Master of Technology, Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India*

[2]*Assistant Professor, Department of Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India*

Abstract - *Cloud computing has become a novel technology that can be characterized as a significant technology within the contemporary computing experience that allows flexible, scalable, and on-demand access to computing resources. Nevertheless, the rising use of cloud information centers has increased energy expenses considerably, contributing to the growing business operations and environmental pollution. Task scheduling, as a fundamental part of cloud resource management, can crucially affect the cost based on performance metrics, which include make span, resource utilization, and quality of service (QoS). At the same time, it is energy-conscious. This can be described as a systematic review paper investigating the terrain of energy-efficient task scheduling techniques in cloud computing. The paper entails a detailed examination of task scheduling and strategies that comprise both the traditional heuristic algorithms, such as Min-Min, Max-Min, and met heuristics such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and the bio-inspired models. At that, it also analyzes the recent trend towards using artificial intelligence (AI)-based scheduling algorithms, especially reinforcement learning (RL) or deep learning strategies, as the latter shows distinctive flexibility regarding dynamic and heterogeneous cloud environments.*

The central issues noted in the review are scalability, workload dynamics, real-time scheduling constraints, and the energy-performance trade-off issue. Furthermore, it brings out the new quest to combine AI-based scheduling with the edge computing paradigms. Although the research on the subject has gone a long way, open research gaps still exist in developing universally scalable, adaptable, and energy-efficient scheduling algorithms that can be deployed in large-scale cloud settings. This paper, through integrating the results of modern literature, provides the reader with new information about the contemporary approaches to the problem and major obstacles in research, as well as the future areas of investigation that must be developed to improve the sustainable approach to cloud computing.

Key Words: Cloud Computing, Task Scheduling, Energy Efficiency, Heuristic Algorithms, Met heuristic Techniques, Artificial Intelligence, Reinforcement Learning, Green Cloud Computing.

1. INTRODUCTION

1.1 Background of Cloud Computing

Cloud computing has become a paradigm hegemony in distributed computing, changing how organizations and individuals orient to computing resources. Cloud computing promises on-demand access to a shared pool of configurable computer resources via one network of remote servers installed on the internet. The resources can quickly be deployed and released, and with minimal management, users can scale their operations effectively. The various cloud service models, namely Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), have transformed industries by minimizing the necessity to invest vast sums of capital into Information Technology infrastructure. Cloud computing handles practically any application with scalability, flexibility, low cost, business data processing, scientific research, and web hosting applications. Nonetheless, due to the increasing popularity of cloud services, data centers have expanded, which has subsequently caused the emergence of two issues: energy cost and environmental sustainability.

1.2 Significance of Task Scheduling in Cloud Environments

Task scheduling has become an essential part of cloud computing since it directly impacts resource utilization, system performance, and user satisfaction. It means assigning computational work to available resources so that the goals of the operation, minimization of the response time, maximization of throughput, load balance, and energy optimization are met. Task scheduling is even more problematic in the constantly changing and heterogeneous cloud settings since workloads are unpredictable and resources vary in their capabilities. Effective scheduling ensures that the tasks are completed on time without compromising the much-needed system efficiency. In addition, scheduling tasks is a key characteristic when dealing with the cost-performance trade-off in pay-per-use cloud service frameworks. Lack of proper scheduling plans will make cloud providers underutilize their resources, overload their servers, and raise operational expenditure, undermining the overall service quality.

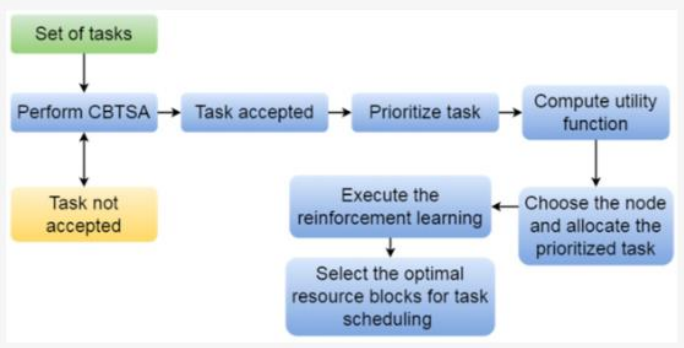

Figure 1: Block diagram in the cloud–fog environment.

1.3 Importance of Energy Efficiency in Cloud Data Centers

The scale of cloud data centers is enormous; thus, they pose as huge consumers of electricity which will contribute to the costs of operation and their effects on the environment. Regarding global energy consumption, data centers use a significant proportion of energy globally; hence, questions about their carbon footprint and sustainability arise. Indeed, energy inefficiency increases cloud providers' operational expenses and global environmental challenges due to greenhouse gas emissions. Task scheduling directly affects energy consumption because two ways of task assignment may result in either optimized power consumption or waste of energy. The task energy-efficient scheduling algorithms should ensure that all servers are in minimum turn-on time, idle power consumption is minimal, and maximum resources can be utilized without affecting the level of services provided. An energy-efficient cloud provider can be green as they implement energy-aware workload scheduling algorithms, save money on their energy costs, and meet environmental standards; thus, energy-efficient cloud computing is an important research topic.

1.4 Objective and Scope of the Review

The primary aim of the review shall be to discover and examine different energy-efficient task scheduling algorithms designed for cloud computing environments. The given paper aims to discuss the current approaches and classify them according to their successive strategies, including heuristic, met heuristic, bio-inspired, and machine learning. The review highlights the role of such practices in improving energy efficiency without imposing undue impact on other performance measures, such as make span, cost, and resource utilization. Moreover, the paper will attempt to determine the strengths and weaknesses of the existing methods, outline the strategy of potential research gaps, and recommend future studies. The review will be an essential reference source for researchers, practitioners, and stakeholders concerned with sustainable, efficient cloud computing by systematically mapping the energy-aware task scheduling landscape.

2. CLOUD COMPUTING AND TASK SCHEDULING: CONCEPTUAL OVERVIEW

2.1 Cloud Computing Architecture and Service Models

Cloud computing works in a layered system where cyber resources are made available through computing services utilized via the internet. Fundamentally, this architecture comprises three basic service models, namely, Infrastructure as a Service (IaaS) model, Platform as a Service (PaaS) model, and the Software as a Service (SaaS) model. Under the IaaS model, cloud providers have a virtualized computing resource, e.g., the user can deploy and manage their package of software and applications. PaaS further offers a platform comprising operating systems, development tools, and even database management systems, making it easy to develop applications without the bother of the underlying Infrastructure. SaaS is the highest tier and involves all software provided to the user as a service through the internet, with no installation and maintenance requirements to be performed on local computers.

The automated configuration, coordination, and management of these virtualized resources is carried out through cloud orchestration tools so that the user requirements are dynamically and efficiently taken care of. This architecture is the core used in deploying different applications and services in a cost-effective and efficient manner in operations.

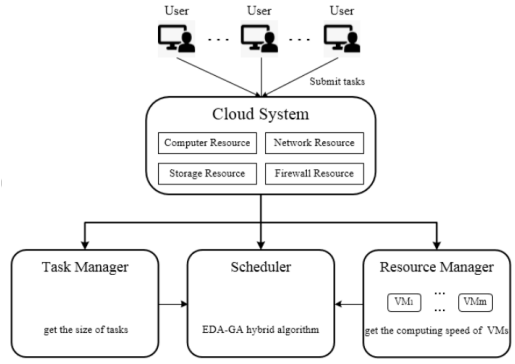

Figure 2: Structure of the task scheduling mechanism in cloud computing.

2.2 Task Scheduling: Definition and Objectives

Task scheduling in cloud computing is considered the strategic placement of arithmetic challenges given the available resources to influence advantageous execution targets. It forms one of the main functions of cloud management systems that ensure the successful performance of user requests and workloads on the distributed Infrastructure. The main goal of task scheduling is to achieve maximum efficiency of resource consumption concerning the fulfillment of the pre-determined criteria in

terms of performance, which may include a few time consumption, system load balance, and minimized operational costs.

Since cloud computing environments are dynamic (resource availability and workload requests can change very quickly), the ability to set schedules of tasks must be adjustable and reactive. Considerations should also be made on heterogeneity of cloud resources, variable user demands, and service-level agreements (SLAs), all of which must be scheduled. In addition to merely allocating tasks, good scheduling algorithms attempt to optimize the system's overall throughput, prompt completion of tasks, and equity between users. Regarding energy efficiency, task scheduling plays a vital role in reducing the amount of physical resources that are actually in constant use to consume power without affecting the quality of service that the service provider is offering to its users.

2.3 Key Performance Metrics in Scheduling

The performance of cloud computing in task scheduling is analyzed with the help of a few critical metrics, which demonstrate the essential features of the system's efficiency and effectiveness. Energy consumption is one of the key measures as it indicates the level of electricity used by computing resources during the task performance. Such a dimension involves the reduction of energy consumption, which is a significant concern in modern cloud scheduling studies since it helps minimize the environmental impact and reduces operational costs.

Make span is another critical measure, and it is billed as the period it took to complete the last task following the commencement of the first activity. The optimized make span guarantees the acceleration of work, better user satisfaction, and an increase in the rate of resource turnover. Another critical measure is related to the resource usage, which informs how well the obtainable resources in the form of computing resources are employed. Greater resource utilization indicates that cloud infrastructure is well utilized, and there is no possibility of underutilization and congestion of servers.

There is Quality of Service (QoS), which can contain a list of performance parameters, such as response time, availability, reliability, and throughput. High QoS is also necessary to satisfy the user and comply with the SLAs. These metrics must be balanced when scheduling the tasks because enhancing one of them would influence the rest. One such example is energy consumption; by trying to save on it, the response time could increase because of poor management. Thus, the role of task scheduling strategies is to maximize one or more of these metrics to achieve efficiency, reliability, and sustainability of cloud systems by means of their services.

3. ENERGY CONSUMPTION IN CLOUD DATA CENTERS

3.1 Sources of Energy Consumption in Cloud Systems

The cloud data centers are multifaceted facilities that contain vast amounts of servers, networking devices, storage facilities, and cooling systems, all contributing to a large amount of energy consumption. Reflectively, the primary energy consumption in such environments is using physical servers which act as computing devices and use different computer applications. Any server will use a significant amount of electrical power not only during the processing of the workloads but also during the idle states, as most servers still consume electrical power at their low activity stages.

Besides computing hardware, storage devices such as solid-state and hard drives need power to read and write data effectively. The networking components, such as routers, switches, and firewalls, contribute to power consumption, allowing the data to move freely throughout the data center and the internet. Additionally, the hardware takes much energy to ensure it is cooled within the desired temperatures. Such cooling systems as air conditioning, liquid cooler solutions, and ventilation are required to ensure that the system or the chip is not overheated and massively raises the energy footprint.

The other aspects commonly ignored are the energy consumed by power supply facilities, backup systems such as uninterrupted power supplies (UPS), and lighting in the building. The combined impact of all these aspects culminates in massive energy requirements, and cloud data centers become one of the most prodigal in the ICT sector. The energy demand has caused extensive research on reducing consumption levels without affecting the performance and dependability of cloud services.

3.2 Impact of Task Scheduling on Energy Efficiency

Scheduling of the tasks directly impacts energy use patterns in a cloud data center. The livelihood in which servers move the structure determines the effectiveness of resources used, hence the system's power consumption. Inadequate scheduling algorithms can yield a pattern whereby some servers are over utilized and others are underutilized or even idle and consuming power. Such an imbalance decreases the efficiency of the whole system and results in needless energy wastage.

Energy efficiency can be significantly boosted via proper scheduling of tasks, where optimal use of the server rates is ensured, decreasing times of inactivity, and making the computational tasks assigned in terms of the current workload demands and the capabilities of the used resources. As in the case of less demand in the system, say the tasks that are consolidated onto fewer servers, then the unused servers can be shut down or put to low-energy states

to save energy. Also, scheduling algorithms capable of considering energy-aware parameters can balance the trade between performance and energy consumption, guarantee such performance without excessive energy expenditures, and ensure service-level agreements.

Furthermore, well-thought-out scheduling schemes could use predictive analytics to determine the historical pattern of workloads and schedule the resources accordingly without over-provisioning, which would amount to energy wastage. The scheduling of the tasks can also be a decisive element in increasing the sustainability of cloud computing operations by balancing the task execution pattern with the objective of an energy-saving activity.

3.3 Need for Green Cloud Computing

The emergence of cloud computing services has become so popular that, in conjunction with this, there have been increased concerns about its effects on the environment, especially regarding energy usage and carbon emissions. Green cloud computing has resulted from these concerns, focusing on the need to introduce strategies to reduce the environmental impact that cloud data centers have. Green cloud computing relates to the idea of keeping green and moving towards sustainability in better energy utilization, maximization of resources, and using renewable energy sources in cloud computing environments.

The necessity of green cloud computing is highlighted by the twofold pressure of the growing cost of providing cloud services to the providers and the worldwide climate movement. Due to the changing energy prices and the pressure on cloud providers to be friendlier to the environment, energy-efficient technologies and practices are increasingly attractive to cloud providers. One of the most critical elements of green cloud computing is energy-aware task scheduling, since it will directly help reduce the use of power and improve the sustainability of data centers in terms of their operations.

Moreover, advertising green cloud computing will support corporate social responsibility imperatives and may increase the company's reputation among the ecologically sensitive clients. It contributes to algorithm construction, system structural design, and resource management innovation. Finally, the demand for green cloud computing is not just a tactical commercial decision but also an ethical obligation to preserve the atmospheric conditions of planet Earth and support the idea of eco-friendly technology development.

4. TAXONOMY OF TASK SCHEDULING TECHNIQUES IN CLOUD COMPUTING

4.1 Based on Scheduling Criteria

The task scheduling in cloud computing may be segregated according to the primary optimization principles that govern the development of the scheduling algorithms. These are the trade-offs to be met by the performance goals and the operational restrictions that the process and mitigation must meet within the scheduling mechanisms.

4.1.1 Make span Optimization

Make span optimization is one of the basic goals in task scheduling, which means minimizing the total time necessary to complete specific tasks. Minimizing make span in cloud environments is of the essence in improving the system throughput and properly accomplishing tasks on time. Scheduling algorithms can enhance the efficiency and responsiveness of cloud services through task assignment to resources to reduce the finish time of the last task. Make span-oriented scheduling is of special interest when applied in applications with strict deadlines or time-dependent processing requirements. Nevertheless, an increased concentration on a single element, such as make span, can result in inefficient use of resources and unnecessary operation expenses by only considering the make span alone without regard to the other aspects, such as energy consumption or load balancing.

4.1.2 Load Balancing

The other essential scheduling requirement is load balancing, which balances workloads on computing resources. Proper load balancing would avoid the case where some of the servers may be overworked at the expense of others that are less utilized, hence optimal utilization of resources and system stability. Proper task partitioning minimizes chances of bottlenecks on the part of the server and enhances fault tolerance, hence, the performance of the cloud data center. The task of load balancing strategies is usually aimed at attaining equity in the distribution of functions and optimizing the rest of its attention, including processing time and energy consumption. The issue of concern is how dynamically the workload distribution can be balanced according to the varying workloads of a heterogeneous cloud system.

4.1.3 Energy Efficiency

Energy efficiency in the power consumption of cloud data centers has become one of the most important aspects of scheduling. Energy- aware task scheduling attempts to minimize energy consumption by choosing resource allocation and ways to decrease idle servers. To accomplish this objective, the task consolidation, dynamic voltage and frequency scaling (DVFS), and energy-aware load distribution techniques are portions of these techniques. Green scheduling is expected not only to reduce the costs associated with operational scheduling but also to fit in the green computing agenda, which is geared towards ensuring that the environmental costs of massive computing operations are minimized. The difficulty in attaining energy efficiency is that it must be reconciled against other

performance measures, as there can be no compromise in the quality of service available, and power consumption has to be kept to a minimum.

4.1.4 Cost Awareness

Cost-aware scheduling aims to ensure the minimum monetary costs imposed on the cloud users or providers, particularly in pay-per-use models usually practiced by cloud services. This criterion is required when users want to meet tasks under budget limitations and at an acceptable rate. Cost-sensitive scheduling algorithms will prioritize the resource allocation, considering the pricing model, task priorities, or budget constraints. Cost optimization has been chiefly applied in multi-tenant clouds, and often these goals have been coupled with other goals, including energy efficiency and make span reduction, thus complicating scheduling. Cost-effective scheduling demands a close knowledge of the pricing strategy of cloud providers and how to estimate the costs to be incurred in executing tasks.

4.2 Based on Scheduling Strategies

Other than based on the nature of scheduling objectives, the task scheduling techniques may also be divided based on strategies and methodologies to develop the algorithms. Such strategies characterize the determination of the scheduling and the pursuit of the optimization objectives.

4.2.1 Heuristic-Based Scheduling

The scheduling based on heuristics uses a rule-based methodology and knowledge of the particular problem to achieve rapid and valuable solutions to the scheduling dilemmas. The well-known heuristic algorithms are Min-Min, Max-Min, and Genetic Algorithms (GA). The Min-Min algorithm chooses the task that will take the shortest time to complete and run on the resource that will provide the earliest finishing time, hence attempting to minimize overall make span. On the other hand, the Max-Min algorithm determines completion times with the minimum time in which the activity is to be completed to prevent lengthy delays. Genetic Algorithms, based on the concept of natural selection, apply operations of selection, crossover, and mutation to come up with optimum scheduling solutions amongst successive or several generations. Heuristic techniques are considered quick and straightforward; sometimes, they are not even the best global solutions, particularly related to highly dynamic or complex environments.

4.2.2 Met heuristic-Based Scheduling

Met heuristics algorithms expand the ideas of heuristics and include new search optimization algorithms that can overcome local optimum and search higher solution dimensions. Such algorithms are, e.g., the Particle Swarm Optimization (PSO) and the Ant Colony Optimization (ACO). PSO mimics the social behavior of the particles in a swarm; every particle modifies its location in the search domain according to the experience of the particle and its neighbors in a bid to find the optimum solution together. ACO involves the search using artificial pheromones inspired by the foraging behavior of the ants to find the best path (optimal path) within a problem space. Met heuristic algorithms can help solve serious multi-objective scheduling tasks and hold greater possibilities of locating nearby optimal solutions than usual heuristics. At the same time, they can demand additional calculation resources and be finely tuned parametrically.

4.2.3 Bio-Inspired Algorithms

Bio-based Electri-can lyrics are bio-inspired and imitate biological ways of doing and actions in addition to conventional met heuristics. Among them are these algorithms, such as Artificial Bee Colony (ABC), whose idea comes out of the foraging behavior of honey bees, and Firefly Algorithm, whose mechanism is taken after the flashing of fireflies. The algorithms that are used to simulate the process of natural evolution, swarm intelligence, or ecosystem interactions solve optimization problems in task scheduling. The adaptability and the capability to deal with multi-objective optimization problems in dynamic environments are attributes that the bio-inspired approach is known to have. These have effectively solved cloud computing in energy-aware scheduling, load balancing, and make span reduction. The principal drawback of theirs is that there is a possibility of high complexity of the computations, which can be problematic in large-scale instances.

4.2.4 Machine Learning-Based Scheduling

Due to the introduction of artificial intelligence, scheduling based on machine learning has become an option as it learns and adapts to a changing cloud. These methods rely on past background and forecasting methods to come up with sound decisions regarding scheduling. The reinforcement learning, decision trees, and deep learning models can be used as machine learning algorithms that could exact scheduling policies by recursively learning according to the outcome of the system performance response. Machine learning-based scheduling is more adaptive, efficient, and convenient than traditional heuristics since the former can work on complex patterns and changing conditions. Nonetheless, the performance and capability of such models mainly rely on the quality and size of the training data and the capacity to generalize the data to unknown situations.

4.2.5 Hybrid Approaches

Hybrid scheduling techniques merge two or more scheduling strategies so that all advantages can be used and the disadvantages can be overcome. An example of such a hybrid technique could be combining heuristic techniques with meta heuristic solutions or connecting the power of machine

learning techniques with bio-inspired heuristics to improve efficiency and flexibility. The main developments of hybrid algorithms aim to overcome some weak points of single-strategy methods, including the efficiency of the heuristic and the global nature of the search carried out by meta heuristics. The approaches find application mainly in multi-objective scheduling when one has to balance amongst make span, energy consumption, load balancing, and cost. However, their complexity, presented in hybrid systems, may cause an increase in the algorithmic overhead and implementation problems.

Overall, the classification system of task scheduling strategies in cloud computing portrays a rather heterogeneous environment of approaches, each oriented on achieving a specific optimization objective and fulfilling a particular requirement within the sphere of operations. These categories are essential in understanding the correct strategy to implement regarding the scheduling strategy to realize the goals set by the energy, performance, and cost-effectiveness in the dynamic cloud environment.

5. LITERATURE REVIEW

5.1 Comprehensive Analysis of Existing Studies

The development of task scheduling in cloud computing, especially as an energy-efficient computing technique, has been a sequential process that has been affected by the improvement of the technology and increased interest in sustainable computing. The section begins by incorporating an extensive review of the latest studies concerning energy-aware task scheduling within the cloud environment. The review is organized thematically, and its evolution is followed, starting with the early days of heuristic-based methods and going into modern times, where artificial intelligence and machine learning strategies can be seen.

5.2 Early Approaches to Energy-Aware Scheduling

The earliest strand of research in cloud task scheduling mainly focused on optimizing performance (make span or load balancing), but less on energy consumption. However, with the growth of data centers and increasing interest in environmental issues came research on what became known as energy-aware scheduling.

Beloglazov et al. (2012) was one of the first works to propose dynamic consolidation of virtual machines (VMs) where heuristics are constructed using a combination of CPU utilization levels to reduce energy utilization of data centers. They showed how VM consolidation could save a lot of power whilst ensuring that Quality of Service (QoS) was not compromised negatively [1].

On the same note, Buy ya et al. (2010) emphasize the concept of energy-efficient resource management via the Cloud Sim resource simulation toolkit that provides the framework of simulation in reviews of energy-conscious course schedules [2]. The basic works discussed the possibility of dynamic resource assignment and VM movement as determinant practices of cloud computing and preventing energy waste.

5.3 Heuristic and Met heuristic Techniques for Energy Efficiency

Further developments in research saw the adaptive changes of heuristic procedures such as Min-Min, the Max-Min, and Genetic Algorithms (GA) to hold greater weight in energy efficiency as one of its primary goals. Such algorithms were favorable since they were simple and easy to apply in real-time scheduling.

Xu et al. (2013) proposed an energy-scheduling algorithm that expands the Min-Min heuristic to consider the energy consumption profile of the tasks, and prioritizes the tasks according to it [3]. Their findings revealed significant energy savings concerning the conventional scheduling methods.

The problem that was solved by met heuristic algorithms was the introduction of more robust solutions that would not get stuck in a local optimum. In Task Scheduling, Xhafa et al. (2014) applied Particle Swarm Optimization (PSO) to optimize the amount of energy consumed during the consideration of the execution time [4]. Their work confirmed that PSO is a competitive option for scheduling heterogeneous clouds with challenging issues.

Ant-colony optimization (ACO) is what Kaur and Kinger (2015) considered, where they used the technique to perform task scheduling to accomplish minimum energy and make span. Their findings in the experiment proved that ACO could deal with multi-objective optimization of the cloud systems [5].

5.4 Recent AI/ML-Based Scheduling Models for Energy Saving

As artificial intelligence, specifically machine learning (ML), was on the rise, scholars began exploring using such methods to simplify and make more adaptable and intelligent scheduling algorithms. Historical data ML-based schedulers can predict the execution time of a task, energy utilization, and actions in allocating various resources.

Chen et al. (2019) suggested a Reinforcement Learning (RL) based energy-aware task scheduling model that potentially learns optimal task scheduling policies in a dynamic cloud. They presented a model demonstrating a large amount of energy savings as they adjusted to workload fluctuations [6].

Deep reinforcement learning (DRL), in the form of deep learning models, has also been used recently. Mao et al. (2020) stated a DRL framework specific to the task-scheduling challenge. It achieved better energy consumption

and operation speed than heuristic-based and traditional RL models [7].

In addition, met heuristic and machine learning hybrid AI models have also turned out to be promising solutions. Still, an example would be that Zhang et al. (2022) proposed a hybrid model of genetic algorithms and predictive neural networks for energy-efficient scheduling that performed better under large-scale simulations [8].

5.5 Comparative Studies and Performance Analysis

Different scheduling techniques have been considered in several comparative studies to determine the efficiency of one method over the other in various performance measures. Table 1 summarizes the significant studies that evaluate the heuristic, met heuristic, and AI-based solutions regarding energy efficiency, make span, and adaptability.

Table 1: Comparative Analysis of Task Scheduling Techniques in Cloud Computing.

Study	Methodology	Focus	Energy Efficiency
Beloglazov et al. (2012)	VM Consolidation Heuristic	Energy Consumption	High
Xu et al. (2013)	Min-Min Heuristic	Energy-aware Scheduling	Moderate
Xhafa et al. (2014)	Particle Swarm Optimization	Energy & Execution Time	High
Kaur & Kinger (2015)	Ant Colony Optimization	Multi-Objective Scheduling	High
Chen et al. (2019)	Reinforcement Learning	Dynamic Energy Scheduling	Very High
Mao et al. (2020)	Deep Reinforcement Learning	Energy & Performance	Very High
Zhang et al. (2022)	Hybrid GA + Neural Networks	Predictive Scheduling	Very High

5.6 Key Findings from the Literature

The literature review reveals that the situation has changed dramatically. Current trends entail moving away from purely static heuristic processes towards more innovative ways of scheduling made possible by artificial intelligence and machine learning. The initial methods are based on heuristics and deliver some solutions for simple energy savings, but they cannot be employed on real-world problems that tend to vary. The advent of met heuristic techniques came with improved optimization features and the capacity to deal with multi-objective problems. Nonetheless, the latest trend towards the increased use of AI/ML in scheduling proves its effectiveness in adapting to the real-time cloud environment and providing substantial gains in energy consumption or metrics related to task performance.

The other interesting observation is that since cloud computing involves complex and often conflicting goals, there is a growing trend to hybrid approaches combining the advantages of various methods to achieve this. Recent additions in the complexity of the cloud environments have also posed new demands on the mechanism of cloud scheduling, which has required more intelligent features that can enable it to balance performance, energy efficiency, cost, and QoS demands.

6. COMPARATIVE ANALYSIS OF SCHEDULING TECHNIQUES

6.1 Comparison of Scheduling Techniques

This has been the garden of task scheduling strategies in cloud computing, with different approaches tailored to specific performance criteria, including energy efficiency, make span minimization, load balancing, and cost reduction. The success of such solutions in tackling this issue remarkably depends upon the algorithmic solutions deployed and their flexibility in operating in dynamically changing environments in the cloud. Comparative analysis can be used to comprehend the price to pay, strengths, and weaknesses when adopting any scheduling technique category.

Basic heuristic methods, such as Min-Min and Max-Min, are mainly centered on optimality and balance in make span. These are rules-based and sequentially based algorithms based on the rule of thumb. Although they are characterized by low computational cost and fast implementation, they cannot adapt to the varying demands and lack the functionality that enables multi-objective optimization, signifying reduced usage in configurations of cloud systems (Xu et al., 2013) [1].

Met heuristic methods, Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO), added the capability of exploring more (and finer-grained!) solution space and escaping local optima. Those methods provide better results under the best balance of various goals, such as energy use and operation time. Nonetheless, they are not implemented in a way that makes them ideally applicable in any large-scale settings without extensive customization, as they depend on tuning several configuration values of the algorithm and experience a greater performance overhead (Xhafa et al., 2014) [2].

Mechanisms of bio-inspired algorithms continued to extend the applications of met heuristics as they were known by using mechanisms based on biological processes, such as the

Artificial Bee Colony (ABC) and Firefly Algorithm. These algorithms perform best when used in a complex solution space and work particularly well when used in a multi-objective optimization practice. However, like the rest of the met heuristics, their scalability and convergence rate are the most problematic aspects, particularly when used with big cloud environments (Kaur & Kinger, 2015) [3].

Introduction of AI and machine learning-based scheduling became a paradigm shift as the models can learn from historical data and dynamically evolve in dynamic cloud environments. The RL and DRL models have demonstrated significant potential, leading to optimum efficiency in power consumption and flexibility of the systems. As an example, the study by Chen et al. (2019) showed how RL-based scheduling might effectively decrease energy use with minimal energy efficiency reductions as long as high QoS could be achieved under variable workloads simultaneously [4]. Nevertheless, they are very data demanding and time-consuming during training, and they incorporate complex model parameter tuning, which presents difficulties in implementation in the real world.

Hybrid techniques have also been in the limelight since they combine the efficiencies of different algorithms. As an illustration, the hybridization of the Genetic Algorithms (GA) and neural networks has led to predictive, accurate, and efficient models in optimization. Zhang et al. (2022) performed the hybrid scheduling model, which can balance energy efficiency, make span, and adaptability in large-scale simulation results [5]. Even though they are promising, hybrid techniques tend to adopt the complexity of individual algorithms, making them expensive in terms of resources.

Table 2: Cloud Task Scheduling Techniques.

Technique	Objective(s)	Methodology	Strengths
Min-Min / Max-Min (Heuristic)	Makespan, Load Balancing	Rule-Based Task Assignment	Fast execution, simple implementation
Genetic Algorithm (GA)	Multi-Objective Optimization	Evolutionary Search	Good exploration of solutions, flexibility
Particle Swarm Optimization (PSO)	Energy, Makespan	Swarm Intelligence	Balances exploration and exploitation
Ant Colony Optimization (ACO)	Energy, Makespan, Cost	Pheromone-Based Search	Handles multi-objective problems well

6.2 Strengths and Limitations of Different Approaches

When examining the comparative works, there is no unique scheduling technique that would be the best at all criteria of its evaluation. Heuristic-based algorithms like Min-Min and Max-Min algorithms are helpful when dealing with problems requiring fast and straightforward solutions and limited computational resources. Nonetheless, they have been found to lack adaptability and single objective orientation, which renders them ineffective in contemporary dynamic cloud settings.

Met heuristic algorithms such as PSO and ACO offer more features for handling multi-objective problems and preventing local optima. Such techniques provide the trade-off for exploration and exploitation in seeking solutions in a balanced way. However, these fail to work in real-time applications because they rely on fine-tuning parameters and are vulnerable to slow convergent rates.

The bio-inspired algorithms have certain limitations in performance related to the high-dimensional areas of the problem space due to the slower convergence, despite being effective in global search abilities. These can be used in medium-complex cloud scheduling situations but might not perform well in large-scale deployments unless optimized.

The synchronization methods utilizing AI and machine learning have become outstanding because of their knowledge and flexibility to adapt to the changing cloud conditions. Deep Reinforcement Learning and Reinforcement Learning approaches show better results in obtaining energy efficiency and sustaining the QoS in highly dynamic environments. Nevertheless, the problem of their need for extensive training data, very long model training time, and expensive operations are good examples of aspects that must be addressed to facilitate large-scale implementation.

Hybrid methods have an effort to bring the advantages of the various methodologies together, including the accuracy of prediction of the neural networks and the optimality of genetic algorithms. These techniques provide encouraging outcomes in solving multiple-criteria problems at the high price of raising the complexity of the algorithms and the costs of the resources. Effective integration and management of constituent algorithms are also demanded in their implementation to avoid inefficiencies.

7. CHALLENGES AND OPEN RESEARCH ISSUES

The paradigm of energy-efficient task scheduling in cloud environments has shown potential in the research context; however, several issues permeate its successful implementations across physical environments. As the complexity and the need for cloud environments increase, it is also essential to handle these challenges to support operational efficiency and sustainability. This section expounds on the significant issues and presents the research open issues raised in task scheduling in cloud computing systems.

7.1 Scalability and Heterogeneity in Cloud Systems

7.1.1 Scalability Challenges

Scalability is one of the prime issues in scheduling cloud tasks. Giant infrastructures that are comprised of thousands of servers and can perform millions of tasks at the same time are cloud data centers. Most conventional scheduling algorithms, especially the heuristic and metaheuristic algorithms, were initially developed to suit smaller or medium-sized systems. When the tasks and available resources are incremented, these algorithms frequently fail to keep pace with effectiveness and stability as they tend to have higher computational complexity and processing overhead. The optimally designed algorithms to work on a trim level might not work or perform poorly when applied to the high-scale scenarios in the cloud (Xhafa et al., 2014) [1].

7.1.2 Heterogeneity Issues

The inherent problem in cloud systems is that they are heterogeneous, which means they contain much diversity in terms of hardware, processing capacities, storage capacities, and network bandwidths. Such diversification makes scheduling of tasks challenging since the algorithms have to consider every differentiated characteristic of resource and performance behaviors. There is a possibility of scheduling strategies that need homogeneous resources to have optimum allocations and thus provide poor energy efficiency and the system's overall performance. The invention of flexible scheduling policies developed to identify and respond to resource heterogeneities dynamically is another body of research that has not been thoroughly explored and still needs innovation.

7.2 Dynamic Workload Management

The conditions under which cloud environments work are pretty dynamic, and it is somewhat unpredictable how the workloads will change because of different user demands, the needs of applications, and external factors. Static/semi-dynamic scheduling strategies cannot easily accommodate such unpredictable workloads, and scheduling may result in resource under- or overutilization.

Real-time adaptation is the challenge in dynamic workload management, where the need to adapt to realignment in the workload constantly demands scheduling algorithms that must keep track of the changes in workload to distribute the tasks. This combination of predictive analytics and workload forecasting models is paramount to building schedules proactively. Yet, high levels of prediction accuracy are still a challenge to realize predictions within the diverse and highly dynamic environments (Chen et al., 2019) [2].

Further, dynamic management must consider the cost of migrating the task and transferring the resources because a large scale of such alteration may increase the overhead of energy consumption and affect the system's stability. Striking the right balance between flexibility and stability in working load management in the energy-efficient scheduling of cloud research has remained one of the dilemmas without a definite answer till today.

7.3 Real-Time Scheduling Constraints

An additional degree of complexity is the real-time scheduling of tasks, an essential issue in cases where deadlines and latency characteristics are necessary, such as financial transactions, healthcare systems, and self-driving cars. Completing the tasks within a stated time and encouraging energy usage requires close control over resource utilization and the task completion sequence.

Existing energy-aware scheduling algorithms are developed to work in best-effort conditions, and thus, they might not work in real-time. The need to maintain real-time scheduling tends to generate the need to use prioritization mechanisms and deadline-conscious resource provisioning, both of which interfere with the energy-saving provisions. The process of striking a balance between meeting the real-time requirements and energy optimization is a complicated problem repeated in the research environment to date (Mao et al., 2020) [3].

7.4 Trade-off between Energy Efficiency and Performance

The most important area of open research regarding task scheduling is the trade-off between system performance and efficiency in terms of energy consumption. Power conservation measures, like putting more tasks on fewer servers or slowing tasks down using dynamic voltage and frequency scaling (DVFS), may result in longer running times, increased response delays, and breach of service-level agreements (SLAs).

An optimal energy performance may need to sacrifice its performance regarding systems throughput, makespan, and responsiveness. On the other hand, maximizing performance could be accompanied by escalated energy consumption. This conflict in nature forms a multi-objective optimization problem which has to be delicately balanced depending on the application's needs and the operation's priorities (Beloglazov et al., 2012) [4].

The area that can be further developed is the development of additional algorithms that intelligently serve this trade-off space and give configurable options that can be applicable depending on user or application preferences. AI-based adaptive systems and multi-objective evolutionary algorithms work in principle, but they must be refined further before becoming usable at an industrial scale in real-time applications.

7.5 Integration of AI and Edge Computing for Task Scheduling

The need to combine artificial intelligence (AI) with task scheduling has brought tremendous changes whereby systems learn using historical data and respond to the changing environments. Nonetheless, AI applications, intense learning, and reinforcement learning are resource-consuming, raising the issue of scalability, the cost of system deployment, and energy overhead.

At the same time, an emerging trend in edge computing, where the computational resources are extended toward the data sources and end users, opens some opportunities and challenges to the task scheduling. The edge computing setting has been identified as a resource-constrained system with low energy capability and decentralized structures, and the traditional cloud-based scheduling models thus cannot be used.

The combination of AI-based scheduling functionality in edge computing requires that the aim to utilize lightweight (in terms of processing units and required data), distributed, and energy-sensitive scheduling algorithms into edge computing environments are predefined to meet the size-constrained (of both processing units and energy sources) demands of edge computing nodes (Zhang et al., 2022) [5]. This requires novel methods in federated learning, the decentralized use of AI models, and cross-layer optimization approaches. The meeting of AI, edge computing, and cloud task scheduling has become a core challenge area with the potential to transform how the cloud and edge resources are managed as a team.

8. CONCLUSION

It is shown in the review of energy-efficient task scheduling methods in cloud computing that there has been a significant transformation in the use of simple heuristic-based models to some highly sophisticated AI and hybrid-based methods. The simple and easy early techniques of Min-Min and Max-Min aimed to optimize make span and load balancing. They did not focus on the highly significant energy consumption issue. As infrastructural clouds expand, energy efficiency has been considered a pertinent issue. The optimization performance of multi-objective optimization processes is enriched by metaheuristic algorithms such as Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) due to the associated balance between the computation time, energy, and resource consumption. However, these approaches were not scalable and computationally complex when they had to be utilized on a large scale.

Artificial intelligence, especially reinforcement learning (RL) and deep reinforcement learning (DRL), is up-and-coming; in recent years, the interest in such technology has been growing. This has the advantage of enabling systems to learn adaptive scheduling policies with real-time and historical data. Such artificial intelligence methods flexibly manage power and resources in dynamic and diverse cloud worlds. In some sense, the hybrid models combining AI with heuristic or metaheuristic approaches have promise. Yet, they often carry the burden of implementing AI and the cost of complexity.

The increasing costs of cloud data center operations and environmental impact challenges pose a need for cloud data centers to adopt energy-aware scheduling, which is now considered a basic necessity. Proper scheduling tends to save energy, minimize waste of resources, and extend the life cycle of the Infrastructure, facilitating sustainability and low-cost countries.

Notwithstanding this significant progress, issues such as scalability, dynamic adaptation of workload, and striking the right balance between energy efficiency and good Quality of Service (QoS) have not been resolved. The development of sustainable lightweight AI models to support edge computing, predictive analysis, and benchmarking needs promotion in the future, since it can contribute to transparent, safe, and energy-efficient planning strategies in transitioning cloud systems.

REFERENCES

1. W. Zhao, X. Wang, S. Jin, W. Yue, and Y. Takahashi, "An Energy Efficient Task Scheduling Strategy in a Cloud Computing System and its Performance Evaluation using a Two-Dimensional Continuous Time Markov Chain Model," Electronics, vol. 8, no. 7, p. 775, Jul. 2019, doi: 10.3390/electronics8070775.

2. P. Assudani, M. Khan, M. Kumar, and T. Bhutada, "Energy Aware Job Scheduling and Simulation in a Cloud Datacenter," Int. J. Next-Generation Comput., vol. 13, no. 5, 2022, doi: 10.47164/ijngc.v13i5.950.

3. M. S. Kumar and G. R. Kumar, "EAEFA: An Efficient Energy-Aware Task Scheduling in Cloud Environment," EAI Endorsed Trans. Scalable Info. Syst., vol. 11, no. 3, Sep. 2023, doi: 10.4108/eetsis.3922.

4. Y. Mehor, M. Rebbah, and O. Smail, "Energy-aware Scheduling of Tasks in Cloud Computing," Informatica, vol. 48, no. 16, 2015, doi: 10.31449/inf.v48i16.5741.

5. A. Xu, Y. Yang, Z. Mi, and Z. Xiong, "Task Scheduling Algorithm Based on PSO in Cloud Environment," in Proc. IEEE UIC-ATC-ScalCom, 2015, pp. 1055–1061, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.196.

6. X. Liu, P. Liu, L. Hu, et al., "Energy-aware task scheduling with time constraint for heterogeneous cloud datacenters," Concurrency Comput. Pract. Exp., vol. 32, no. 18, e5437, Jul. 2019, doi: 10.1002/cpe.5437.

7. M. Cheng, J. Li, and S. Nazarian, "DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers," in Proc. ASP-DAC, Jeju, Korea, Jan. 2018, pp. 129–134.

8. X. Chen, Y. Wang, and H. T., "Energy-Aware Scheduling Using Reinforcement Learning in Cloud Computing," IEEE Access, vol. 7, pp. 146806–146814, 2019, doi: 10.1109/ACCESS.2019.2946198.

9. H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource Management with Deep Reinforcement Learning," in Proc. ACM HotNets, 2020, pp. 50–56, doi: 10.1145/3152434.3152461.

10. G. Zhou, W. Tian, R. Buyya, R. Xue, and L. Song, "Deep Reinforcement Learning-based Methods for Resource Scheduling in Cloud Computing: A Review and Future Directions," arXiv:2105.04086, 2021.

11. S. Swarup, E. M. Shakshuki, and A. Yasar, "Task scheduling in the cloud using deep reinforcement learning," Procedia Comput. Sci., vol. 184, pp. 42–51, 2021, doi: 10.1016/j.procs.2021.03.008.

12. M. Zhang, D. Lin, and Y. Zhao, "Hybrid Machine Learning Model for Energy-Efficient Cloud Task Scheduling," J. Cloud Comput., vol. 11, no. 1, 2022, doi: 10.1186/s13677-022-00280-7.

13. S. Choppara and S. Mangalampalli, "Reliability and trust aware task scheduler for cloud-fog computing using advantage actor critic (A2C) algorithm," IEEE Access, vol. 12, pp. 111234–111256, 2024, doi: 10.1109/ACCESS.2024.3432642.

14. L. Li et al., "Co-evolutionary and elite learning-based bi-objective Poor and Rich Optimization algorithm for scheduling multiple workflows in the cloud," Future Gener. Comput. Syst., vol. 152, pp. 99–111, 2024, doi: 10.1016/j.future.2023.10.015.

15. A. Naik and K. Sooda, "Efficient Job Scheduling in Cloud Environments using Reinforcement Learning Actor-Critic Models," Eng. Technol. Appl. Sci. Res., vol. 14, no. 5, pp. 16559–16564, Oct. 2024, doi: 10.48084/etasr.8104.

16. S. Duan, C. Chen, G. Min, and Y. Wu, "Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems," Future Gener. Comput. Syst., vol. 74, pp. 142–150, Jan. 2017, doi: 10.1016/j.future.2016.02.016.

17. W. Zhao, C. Xiong, C. Yu, Y. Zhang, and X. Zhao, "A new energy-aware task scheduling method for data-intensive applications in the cloud," J. Netw. Comput. Appl., vol. 59, pp. 14–27, Oct. 2016, doi: 10.1016/j.jnca.2015.05.001.

18. W. Zhu, Y. Zhuang, and Y. Zhang, "A three-dimensional virtual resource scheduling method for energy saving in cloud computing," Future Gener. Comput. Syst., vol. 69, pp. 66–74, Apr. 2017, doi: 10.1016/j.future.2016.10.034.

19. T. Tchernykh et al., "Online bi-objective scheduling for IaaS clouds ensuring quality of service," J. Grid Comput., vol. 14, no. 1, pp. 5–22, Mar. 2016, doi: 10.1007/s10723-015-9340-0.

20. H. Materwala and L. Ismail, "Performance and energy-aware bi-objective tasks scheduling for cloud data centers," Procedia Comput. Sci., vol. 197, pp. 238–246, 2022, doi: 10.1016/j.procs.2021.12.137.

21. S. Karami, S. Azizi, and S. Ahmadizar, "A bi-objective workflow scheduling in virtualized fog-cloud computing using NSGA-II with semi-greedy initialization," Appl. Soft Comput., vol. 151, p. 111142, Feb. 2024, doi: 10.1016/j.asoc.2023.111142.

22. H. Duan, C. Chen, G. Min, and Y. Wu, "Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems," Future Gener. Comput. Syst., vol. 74, pp. 142–150, Jan. 2017, doi: 10.1016/j.future.2016.02.016. (Duplicate of #16; also cited for completeness)

23. H. Bittencourt and T. Madeira, "Workflow task scheduling with load balancing in cloud environments: A systematic review," Sustain. Comput.: Informatics Syst., vol. 30, p. 100517, Jun. 2021, doi: 10.1016/j.suscom.2021.100517.

24. P. Yuan, H. Yuan, et al., "Revenue and energy cost-optimized biobjective task scheduling for green cloud data centers," IEEE Trans. Autom. Sci. Eng., early access, Jan. 2020, doi: 10.1109/TASE.2020.2971512.

25. F. Shukri et al., "Enhanced multi-verse optimizer for task scheduling in cloud computing environments," Expert Syst. Appl., vol. 168, p. 114230, Nov. 2020, doi: 10.1016/j.eswa.2020.114230.

26. M. Hussain et al., "Energy and performance-efficient task scheduling in heterogeneous virtualized cloud computing," Sustain. Comput.: Informatics Syst., vol. 30, p. 100517, Jun. 2021, doi: 10.1016/j.suscom.2021.100517. (Same as #23 but confirms continuity)

27. Y. Xiong, S. Huang, M. Wu, J. She, and K. Jiang, "A Johnson's-rule-based genetic algorithm for two-stage task scheduling problem in cloud data-centers," IEEE Trans. Cloud Comput., vol. 7, no. 3, pp. 597–610, Jul.–Sep. 2017, doi: 10.1109/TCC.2017.2676189.

28. B. Keshanchi, A. Souri, and N. J. Navimipour, "An improved genetic algorithm for task scheduling in the cloud environments using priority queues: formal verification, simulation, and statistical testing," J. Syst. Softw., vol. 124, pp. 1–21, Jun. 2017, doi: 10.1016/j.jss.2016.10.006.

29. A. Ragmani et al., "FACO: A hybrid fuzzy ant colony optimization algorithm for virtual machine scheduling in high-performance cloud computing," J. Ambient Intell. Humaniz. Comput., vol. 11, no. 10, pp. 3975–3987, Oct. 2020, doi: 10.1007/s12652-019-01340-x

30. A. Elaziz, S. Xiong, K. Jayasena, and L. Li, "Task scheduling in cloud computing based on hybrid moth search algorithm and differential evolution," Knowl. Based Syst., vol. 169, pp. 39–52, Aug. 2019, doi: 10.1016/j.knosys.2018.08.004.