Task Scheduling Techniques for Energy Efficiency in the Cloud

Sanna Mehraj Kak¹, Parul Agarwal¹,* and M. Afshar Alam¹

¹Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India

Abstract

Energy efficiency is a key goal in cloud datacentre since it saves money and complies with green computing standards. When energy efficiency is taken into account, task scheduling becomes much more complicated and crucial. Execution overhead and scalability are major concerns in current research on energy-efficient task scheduling. Machine learning has been widely utilized to solve the problem of energy-efficient task scheduling, however, it is usually used to anticipate resource usage rather than selecting the schedule. The bulk of machine learning approaches are used to anticipate resource consumption, and heuristic or metaheuristic algorithms utilize these predictions to choose which computer resource should be assigned to a certain activity. As per the knowledge and research, none of the algorithms have independently used machine learning to make an energy-efficient scheduling decision. Heuristic or meta-heuristic approaches, as well as approximation algorithms, are frequently used to solve NP-complete problems. In this paper, we discuss various studies that have been used to solve the problem of task scheduling which belongs to a class of NP-hard. We have proposed a model to achieve the objective of reduced energy consumption and CO₂ emission in a cloud environment. In the future, the model shall be implemented in MATLAB and would be assessed on various parameters like makespan, execution time, resource utilization, QoS, Energy utilization, etc.

Keywords: CDC (Cloud Datacentre), CSP (Cloud Service Provider), DA (Dragonfly Algorithm), EDA-GA (Estimation of Distribution Algorithm and GA), FF (Firefly), GA (Genetic Algorithm), IaaS (Infrastructure-as-a-Service), MGWO (Modified Mean Grey Wolf Optimization Algorithm), PaaS (Platform-as-a-Service), SaaS (Software-as-a-Service), SAW (Sample Additive Weighting), SLA-LB (Service Level Agreement Based Load Balancing), TBTS (Threshold Based Task Scheduling Algorithm), TS (Task Scheduling)

Received on 20 January 2022, accepted on 07 June 2022, published on 20 June 2022

Copyright © 2022 Sanna Mehraj Kak et al., licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution license, which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/ew.v9i39.1509

1. Introduction

Cloud computing is used by almost every online user, either directly or indirectly. Cloud computing is a novel computing architecture that makes computing a scalable and dependable utility by allowing us to borrow hardware and software from somewhere on the Internet. CSPs provide and host these software and hardware (e.g. Amazon EC2, Google Compute Engine, and Oracle Cloud) [1, 2]. The CSP provides a variety of services, including IaaS, PaaS, and SaaS. Typically, a cloud data center with more servers is used. Cloud computing has several advantages in terms of total cost and data usability. The cloud provider handles data values from data owners [3, 4]. Furthermore, in the single cloud, there are concerns about service degradation and hostile insiders [5]. The multi-cloud environment, on the other hand, keeps track of several cloud infrastructures and prevents reliance on others [6, 7, 8]. The multi-cloud resource sharing system is based on the storage and device partitioning mechanism [9]. Four types of architecture are utilized to supply services to various cloud vendors. They are application duplication, application device partitioning, fragmented application logic, and fragmented application data [10, 11, 12].

The data center in the cloud handles massive demands for apps from all around the world [13, 14]. There are numerous dependent and autonomous activities in each application. Each cloud requires a scheduling strategy to

*Corresponding author. Email: Pagarwal@jamiahamdard.ac.in
carry out such requirements and determine the task execution order. As cloud computing applications become more popular and in demand, an accumulation of high-performance cloud resources is desirable. The energy consumption rate in the CDC skyrockets as the demand for high-performance computing resources rises. Data centers are expected to consume approximately 1.4 percent of all electrical energy consumed globally, with this figure increasing at a rate of 12 percent per year [15]. “Processing units, cooling facilities, and storage facilities” require the most energy, accounting for 42.0 percent, 15.4 percent, and 14.3%, respectively [16]. Different clouds may have different scheduling algorithms [17, 18].

Figure 1 shows the process of task scheduling in a cloud system. Users transmit the tasks to the cloud, where they are assigned to a processor. It’s now an issue of allocating processor duties in such a way that the CSP obtains the most advantage in the least amount of time, thus maintaining the maximum resource utilization with a minimum makespan. Task scheduling objectives [18, 19] include minimizing the execution time of the task and energy consumption, along with improving the utilization of resources and the capacity to balance loads. With the rapid growth of cloud users, decreasing the time for completion of the task is also useful for improving customer experience [19, 20]. Load balancing capabilities are strengthened in order to avoid performance loss owing to resource expansion or unused resources caused by unneeded futile resources, resulting in maximum VM utilization. As a result, the TS method was discovered to be NP-complete, meaning that finding and concluding the best answer in a finite period is impossible [21, 22, 12]. Green characteristics of brands and facilities have been observed in ICTs in ideas connected to sustainability, such as sustainable computing, green computing, green ICTs, and biodiversity informatics [23].

To date, evolutionary algorithms such as GA and distribution estimation algorithms have been introduced to solve multiple scheduling and mapping problems [24, 25, 26, 27, 1]. Scheduling algorithms that include the Min-Min algorithm, Max-Min algorithm, and Suffrage models implemented with the single-objective model. However, they have poor adaptability and extensibility. Most of the time, the model only takes into account the minimum amount of time needed to finish the work, not the safety measure [28].

The major contributions of this paper are:

- We summarize the related work for the techniques of existing energy-efficient task scheduling mechanisms.
- We plan to develop a novel hybrid optimization-based task scheduling technique that is energy proficient.

This paper is organized as follows: In section 1, a brief introduction to the importance of employing green cloud computing is given. Section 2 explains various problems that need attention while using cloud computing. In section 3, a review of various energy-efficient task scheduling methods in a cloud environment is presented. A comparison of various works’ methodologies, futures, and final outcomes is also provided. We have also formulated a table that includes the methodologies, features of their work, and the final outcomes. In section 4, a model has been proposed where we will be aiming to reduce the consumption of energy in cloud. In section 5, we draw the conclusion; also, a future prospect of the work is explained.

1.2. Problem Statement

Table 1 demonstrates the analysis of task scheduling in a multi-cloud atmosphere. Initially, the Min-Min and Max-Min models were used in [15], which provides minimal complexity with high utilization of the cloud; nevertheless, it needs deliberation on fault tolerance. Also, the SLA-LB model was employed in [16], which provides minimal execution time with a reduced penalty cost. However, it requires consideration of energy utilization. Likewise, EDA-GA was presented in [9], which enhances the load balance and it concerns the minimal execution time. However, it needs consideration in real cloud environments. In addition, the sample additive weighting (SAW) model was deployed in [29], which offers high reliability with reduced makespan. Nevertheless, the continuous arrival of tasks is not considered. The CSSA algorithm was deployed in [30] that is cost-effective and also increased the throughput. However, analysis in real-time has to be concentrated more. A NN model was presented in [31], which had a shorter development time and higher reliability. However, machine learning models are not taken into account. In [32], an ADA model was developed that offers low execution costs with improved load balancing, but it requires more attention to VM scheduling. In [33], an MGWO model was deployed, which provides a reduced makespan with minimal energy utilization; however, it must focus more on task priority.

GCC is a promising and popular field of study because there is a need for stable and long-lasting resources as well as...
as their preservation, which can be accomplished by enhancing the usage of computational services from multiple viewpoints, such as software, network, and hardware optimization methods [34]. When the emphasis is placed on energy conservation at multiple levels of the IoT, it leads to Green IoT (GloT), which ultimately reduces energy costs while keeping human health hazards in mind. The use of energy may be greatly reduced if a data centre is run using a balanced mix of grid computing and sustainable energy [35].

2. Related Work

In 2019, the authors in [15] have created a multi-cloud network in which numerous clouds are linked together to provide a combined service. On the other hand, task scheduling in a multi-cloud was far more difficult than in a single cloud. Three “allocation aware task scheduling algorithms” for a multi-cloud network were proposed in this paper. The proposed model is based on the well-known Min-Min and Max-Min algorithms. Finally, the developed model was compared to traditional models in terms of throughput, average cloud consumption, and makespan.

In 2020, the authors in [16] explored two allocation models termed TBTS and SLA-LB models. They explored two allocation models termed TBTS and SLA-LB models. TBTS schedules the tasks in batches and aids in the scheduling of tasks in virtual machines with diverse configurations. SLA-LB scheduled the tasks in a dynamic manner depending on the necessity of users, like budget and deadline. Their experimental outcomes revealed the performance of the developed model in terms of utilization factor, penalty cost, makespan, and completion time.

In 2019, the authors in [9] constructed a hybrid task scheduling method that was based on EDA and GA. EDA’s sampling and probability models were initially used to provide specific viable solutions. The optimal scheduling strategy for allocating jobs to VMs was discovered in the final phase. Stronger searching capability and faster convergence speed are two advantages of this architecture. The investigational results have illustrated that the presented scheme has efficiently improved the load balancing capability and reduced completion time.

In 2017, the authors in [28] developed a multi-task scheduling model that takes into account the service quantity and service coefficient and incorporates a task workload model. Following that, the influence of several workload-oriented task scheduling approaches on utilization and total completion time was investigated. The scenarios with and without time parameters were thoroughly examined independently. Finally, the simulation results showed that high-priority jobs with a substantial burden decreased the time to completion and enhanced utilization without compromising quality.

In 2019, the authors in [29] explored CSSA for optimal “multitask scheduling in an IaaS cloud environment”. The method developed job plans on a constant basis, making the current schemes more profitable. The prior network was produced with chaotic optimization for well-organized task allocation to provide improved global convergence. In the end, the established CSSA model has revealed better outcomes when compared to the traditional SSA model. In 2020, the authors in [30] investigated a prediction-oriented approach known as the "NN-DNSGA-II algorithm," which combines the NSGA-II and ANN frameworks. In addition, five leading non-prediction-oriented dynamic techniques for workflow scheduling were used to solve the problem. The presented work concerns six objectives, namely, enhancement of utilization and reliability, reduction of energy, cost, makespan, and level of disproportion.

In 2020, the authors in [31] developed a new load balancing task scheduling model in a cloud employing ADA, which was an amalgamation of DA and FF algorithms. Furthermore, the multi-objective function was constructed based on three constraints: load, processing costs, and completion time, in order to achieve higher performance. Finally, the performance of the proposed system was calculated using a variety of metrics such as execution time and cost.

In 2019, the author in [32] introduced a new TS model, with the primary goal of improving the TS by lowering energy consumption and MS. MGO was able to overcome scheduling issues and improve system performance. Encircling and hunting were adjusted in MGWO using the mean value to improve GWO’s efficiency. Finally, the findings showed that the suggested method effectively decreased energy usage when compared to the traditional GWO model.

<table>
<thead>
<tr>
<th>Author (Citation)</th>
<th>Adopted Methodology</th>
<th>Features</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panda et al. [15]</td>
<td>Min-Min and Max-Min model</td>
<td>❖ Minimum complexity</td>
<td>❖ No consideration on fault tolerance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>❖ High utilization of cloud</td>
<td></td>
</tr>
<tr>
<td>Lavanya et al. [16]</td>
<td>SLA-LB model</td>
<td>❖ Minimal execution time</td>
<td>❖ Energy utilization is not taken into account.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>❖ Reduced penalty cost</td>
<td></td>
</tr>
<tr>
<td>Pang et al. [9]</td>
<td>EDA-GA</td>
<td>❖ Improved load balance</td>
<td>❖ Need consideration on real</td>
</tr>
</tbody>
</table>

Table 1. An overview of the related work
3. Proposed Method

Resource allocation in cloud data centers faces considerable problems in terms of dynamism and complexity due to constantly changing system states and varying user demands. Although there are solutions that focus on this issue, they are unable to respond effectively to dynamic changes in system states and user requests because they rely on existing system information [34, 12]. The increased usage of datacenter cloud computing systems is raising concerns about rising power consumption and associated carbon emissions, which will be produced by massive datacenters being built at breakneck speed. It has been estimated that if we keep exhausting our resources, carbon emissions will quadruple in the next 20 years.

To cope with this challenge, we intend to introduce a new hybrid optimization-based energy-efficient task scheduling approach in the cloud computing environment. The proposed work concerns primarily on optimally allocating the tasks via the proposed hybrid optimization concept. Accordingly, the proposed optimal task allocation considers the objectives such as makespan, energy consumption, maximum resource utilization, execution time, and response time. To solve this optimization issue, a new hybrid algorithm will be formulated by hybridizing the concepts of several standard optimization models like sparrow search algorithm (SSA) [36], Deer Hunting Optimization Algorithm (DHOA) [37], Particle Swarm Optimization Algorithm (PSO) [38], and Firefly Algorithm (FF) [39]. The sparrow search algorithm (SSA) was hypothesized and motivated by the collective wisdom, foraging, and anti-predation activities of sparrows. Human hunting behavior toward deer influenced the development of an innovative algorithm, DHOA. Even though the hunters’ actions differ, the manner in which they assault the buck or deer is determined by the hunting plan they devise. Particle swarm optimization (PSO) is a population-based optimization approach influenced by the motion of bird flocks and schooling fish. PSO and the evolutionary computation method have many similarities. The network is started with such a population of random solutions, and the quest for the best solution is conducted via updating generations. The Firefly algorithm (FA) is a dynamic random search method that can demonstrate significant flexibility and robustness for solutions in a dynamic context.

The working strategy of the proposed work is represented in Figure 2.

(a) Makespan: We describe makespan as the time required from when a user provides his inquiry to when the final task unit is completed. It takes into account both processing and waiting time.

(b) Energy Efficiency: Sum of the energy consumption of each unit involved in the service of user requests.

(c) Execution Time: It is the amount of time necessary to complete the task.

(d) Response time: The time taken to respond to a task.

4. Discussion

We are aiming to develop an algorithm that will be used to decrease the consumption of energy, hence reducing the carbon footprint so that our future generations are left with both renewable and non-renewable resources for their best survival.
The creation of generic or standardized rules that cloud service providers must adhere to in order to accurately portray the services they offer is an absolute necessity. At the moment, different manufacturers comply with complex patterns of data types and semantic models, which in turn may result in a conflict of interest being created. Systematic studies need to be undertaken if one wants to achieve simplification in the deployment of cloud services and a seamless deployment [40]. Nowadays, research is going on fog computing, which is more proficient, and provides more security and adaptability. Cloud computing technology has advanced to the point where several development tools for designing and implementing cloud architectures are now accessible [41]. Fog computing is still in its early stages of development, with prototype models and development tools still under development, but it is thought to represent the future of current computing technology, evolving quickly and using the edge of devices for computational purposes [42].

5. Conclusion and Expected outcome

Among the most serious concerns discussed in the CDC is energy-efficient job scheduling. Indirectly, lowering energy usage and makespan lowers administration expenses and enhances the performance of the cloud. It also minimizes the level of CO2 emitted by CDCs. Numerous academics are focused on this critical topic, and various cognitive, hybrid, and machine learning-based techniques have been presented. The difficulty with these techniques is that when the cloud environment expands, they have the tendency to increase the processing time, and they frequently perform duplicate solution searches for the same problem instance. The proposed work will be implemented in MATLAB and the corresponding outcomes acquired will be noted. The performance analysis will be done by comparing the proposed model with several state-of-the-art models in terms of convergence, energy, and makespan, respectively.

References


Figure 2. Working Principle of Proposed Model
Computational Intelligence and Neuroscience, vol. 2019, Article ID 9293617, 21 pages, 6 December 2019.