Artificial intelligence-based cloud-internet of things resource management for energy conservation

Soukaina Ouhame¹, Moulay Youssef Hadi², Amine Mrhari¹, Imane Laassar¹

¹Department of Computer Science, Faculty of Science, Ibn Tofail University, Kénitra, Morocco
²Department of Computer Science, Higher School of Technology, Ibn Tofail University, Kénitra, Morocco

ABSTRACT

The widespread demand for hosting application services in the cloud has been fueled by the deployment of cloud data centers (CDCs) on a global scale. Furthermore, modern apps’ resource needs have sharply increased, especially in industries that use a lot of data. As a result, more cloud servers have been made available, resulting in higher energy usage and, ecological problems. Large-scale data centers have been developed as a result of the rapidly increasing demand for cloud services, allowing application service providers to rent data center space for application deployment by user-required quality of service (QoS). These data centers use a lot of electricity, which raises running expenses and produces more carbon dioxide (CO2) emissions. Modern cloud computing environments must also provide QoS for their users, necessitating a trade-off between power performance, energy consumption, and service-level agreement (SLA) compliance. We present an intelligent resource management policy using enforcement learning for CDCs. The objective is to continuously consolidate and dynamically allocate virtual machines (VMs). Utilizing live migration and disabling inactive nodes to reduce power consumption in this cloud environment while maintaining service quality. To enable dynamic resource management, a better power-performance tradeoff, and significantly lower energy consumption, we integrate several artificial intelligence concepts. Based on the result the proposed approach is more efficient as compared with other techniques.

Keywords:
Artificial Intelligence
Cloud computing
Service-level agreement
Trade-off

This is an open access article under the CC BY-SA license.

Corresponding Author:
Soukaina Ouhame
Department of Computer Science, Faculty of Science, Ibn Tofail University
Av. de L’Université, Kénitra, Morocco
Email: soukaina.ouhame@uit.ac.ma

1. INTRODUCTION

The definition and development of cloud computing have been the subject of intensive research during the last two decades. Cloud computing, which is fueled by advancements in networking and distributed architectures, is a manifestation of distributed systems research that dates back to the inception of the client-server model in 1958 [1]. Because of the cloud's explosive growth, it has become a vital tool in every sphere of society, including government, business, and education. The development of new technologies and paradigms to meet the demands of emerging applications, such as scientific, healthcare, agricultural, smart city, and traffic management, has been made possible by characteristics of cloud computing, such as dynamic, metered access to shared pools of computing resources [2]. To meet heterogeneous quality of service (QoS) needs, well-known cloud providers such as Facebook, Google, and Amazon currently use large-scale cloud data centers (CDCs). A unified interface over heterogeneous
resources found in the internet of things (IoT) can also be provided by cloud computing platforms. Preprint submitted to IoT applications based on the IoT that increases the dependability of cloud services [3]. The size of data in many fields, including scientific computing, data processing, bioinformatics, and IoT activities, has increased due to the rapid growth of distributed cloud computing network services. Thousands of high-performance computers are deployed in CDCs, where they conduct millions of processes [4]. Virtualization is one of the most beneficial advantages for customers who may use the many different kinds of services that the cloud provides via virtual machines (VMs), which are used to provide a variety of services. Typically, these VMs use a lot of energy. Such energy use boosted power prices and had a detrimental effect on the environment [5]. Energy efficiency is a major problem for both customers and providers of cloud services. To do that, researchers present different kinds of researchers in this field using different techniques some of which are as follows.

2. RELATED WORK

Diamanti [6] examined the problem of energy-efficient scheduling in heterogeneous computing. The goal was to use the least amount of electricity possible while scheduling all tasks by their QoS requirements. According to Shuaib et al. [7] illustrate that hybrid data centers built with more energy-efficient servers may reduce electricity consumption. The time of the activity's processing is essentially inextricably linked to the issue of energy consumption. Determining the ideal operating time for activities is complicated by this. Academics and businesses alike are interested in the energy savings that virtualized data centers can provide. Numerous studies on efficient scheduling techniques and multi-component systems have been conducted [8]. Utilizing the intra-task dynamic voltage and frequency scaling (DVFS) scheduling technique to more effectively optimize the energy for real-time activities when systems require more power, based on job profile information, under the assumption of zero transmission delay, described as an integer linear programming (ILP) model. Finding the best processing frequency for each basic block to produce the lowest average energy is the main goal of this ILP model. The ILP algorithm is changed to determine the best execution frequency as well as the proper program locations to insert the conversion instructions to account for the DVFS conversion overhead. To address the issues with work consolidation and scheduling, the author proposed an energy-aware task scheduling algorithm (ETSA) [9]. The suggested ETSA approach enhances task completion time for scheduling decisions and total resource use, and it suggests a normalization technique. The size of data in many fields, including scientific computing, data processing, bioinformatics, and IoT activities, has increased due to the rapid growth of distributed cloud computing network services. Thousands of high-performance computers are deployed in CDCs, where they conduct millions of processes [10–12].

Virtualization is one of the most beneficial advantages for customers who may use the many different kinds of services that the cloud provides via VMs, which are used to provide a variety of services. Typically, these VMs use a lot of energy such energy use boosts power prices and has a detrimental effect on the environment [13]. Energy efficiency is a major problem for both customers and providers of cloud services when IoT is attached. To do that, researchers tried to offer suggestions for how to fix this problem. Some of the most significant research articles related to this issue are presented in this section. Despite numerous attempts to reduce energy usage by using different technologies, it is estimated that the data center releases 62 million tons of carbon dioxide (CO2) into the environment [14]. Several independent tasks were examined in this study utilizing QoS standards. Consequently, choosing the VMs from a resource arrangement with a heterogeneous configuration that uses the least amount of energy to suitable servers to have the best energy use. These tasks were created with the help of numerous researchers. For instance, one study [15] examined the problem of energy-efficient scheduling in heterogeneous devices with IoT and cloud computing. The goal was to use the least amount of electricity possible while scheduling all tasks by their QoS requirements. Research by Jeyaraj et al. [16] illustrates that hybrid data centers built with more energy-efficient servers may reduce electricity consumption. The time of the activity's processing is essentially inextricably linked to the issue of energy consumption. Determining the ideal operating time for activities is complicated by this. Academics and businesses alike are interested in the energy savings that virtualized data centers can provide. Numerous studies have been conducted on multi-component systems and efficient scheduling algorithms to better optimize the energy for real-time activities as systems require more power. According to Jangra and Mangla [17] scheduling the suggested ETSA strategy optimizes task completion time and overall resource use and suggests a normalization method when it comes to scheduling. A proper schedule might not always be produced using conventional methods. Most studies focus on heuristic algorithms, which frequently use greedy local optimal selection heuristics as their foundation. However, several well-known metaheuristic techniques, including ant colony optimization (ACO), particle swarm optimization (PSO), and genetic algorithms (GA), have demonstrated prominence in task scheduling
problems due to their adaptability. A two-stage energy and performance-efficient task scheduling technique (EPETS) was proposed by Hessen et al. [18]. While ignoring energy usage, the initial stage of scheduling helped reduce processing time and satisfy task deadlines. Locating these locations was the second stage of work assignment scheduling. The most effective execution location uses the least amount of energy and does so within the time constraints. The study of energy-saving techniques for VMs configuration, migration, and consolidation has recently grown in importance as a research topic. In reality, several evolutionary algorithm-based methods for resolving the VMs consolidation problem have been proposed. In this context, Hakiri et al. [19] proposed a VMs placement approach that assesses resources, VMs state, QoS measures, and I/O data using a priority-based probability scheduling model. The VMs placement process evaluates data location to reduce unnecessary [20].

3. PROPOSED ALGORITHM

By packing more VMs onto a smaller server and using live migration to reduce resource utilization while maintaining user-specified service-level agreements (SLA), VMs consolidation improves resource management and data center optimization. In the machine learning (ML) field, several learning strategies have been found and proposed over the years. However, since each learning algorithm is tailored to a particular problem, there isn't a single, all-encompassing solution. Therefore, a particular system method must be chosen based on how well it fits the situation [21], [22]. In this research, we present a self-optimizing reinforcement learning (RL)-based VMs migration approach that guarantees increasing data center service by maximizing VMs allocation and obtaining more energy gains. We employ Q-learning. In the ML field, several learning strategies have been found and proposed over the years. However, since each learning algorithm is tailored to a particular problem, there isn't a single, all-encompassing solution. Therefore, a particular system method must be chosen based on how well it fits the situation. In this research, we present a self-optimizing RL-based VMs migration approach that guarantees increasing data center service by maximizing VMs allocation and obtaining more energy gains. We resolve the VMs consolidation issue using the Q-learning approach, and we evaluate the method's success using a range of cloud performance indicators.

The RL agent learns the best resource allocation method based on its frequent interactions with the environment and the current condition of the system. RL enables an agent to learn the optimal activity in a stochastic, nondeterministic environment. As demonstrated in Figures 1 and 2, when an agent interacts with its environment to learn how to improve its behavior and determine the optimal policy to achieve its goals, a cyclical learning of state-action-reward interactions takes place. The agent must experiment to find the actions that reward him the most. Additionally, the chosen action affects both the immediate reward and following states, and consequently all subsequent rewards. The Markov decision process (MDP), which describes learning in sequential decision-making settings with adverse uncertainty, is usually considered the gold standard. The MDP model enables agents to acquire an ideal strategy using simulated [23], [24]. According to Alam et al. [25] this trait, all that is required to predict future states is the environment's current condition. States, actions, transition probabilities, and rewards are all included in the MDP framework.

Figure 1. Illustration of the agent in the interaction system in ClodDC

Figure 2. Illustration of the agent section

States, actions, transition probabilities, and rewards are all included in the MDP framework, denoted by the letters \( S, A, P, \) and \( Q \) respectively \( S = \) Present the set of all possible state; \( A = \) Signifies a set of
actions; $P = (St, at)/st + 1$ represents the probability distribution that controls state transitions; and $Q = (St, at)/st + 1$ is a symbol for the probability distribution governing the rewards received $R = (St, at)$.

The teaching strategy is broken down into distinct periods. At the end of each time step $T$, the learning agent is in the state $S$. The agent selects an action at " $A(st)$ " where " $A(st)$ " denotes the group of feasible actions in the present state. The agent picks a course of action at " $A(st)$ " where " $A(st)$ " refers to the assortment of feasible activities under the existing circumstances $S$. the agent is rewarded $R(st + at_1)$ for doing the required action, which results in a change in the environment $st$. The agent is in state $S$ so chooses action $at_1$, the state transition probability $(St, at)/st + 1$ determines the probability of a shift to $st + 1$ the anticipated compensation that the agent would receive after moving from state $st$ to state $s$ by carrying out $q (= (St, at)/st + 1)$ specifies the anticipated reward that the agent will obtain after changing from state $st$ to state $S$ by carrying out action $out$ [26]. The update rule is defined as $Q(st, at) ← Q(st, at) + \alpha[r + \gamma \max Q(st + 1, s) − Q(st, a)]$ [27].

RL (Q-learning) is the foundation of the suggested solution. This approach enables the learning of a policy based on the actions taken in each environmental state. For that purpose, we implement for VMS migration role using the proposed algorithm. Until every VMs in the migration list has been transferred to a new host in the environment, this process is repeated. Using our best energy management algorithm, we rank all VMs according to their current usage before allocating them to hosts with slightly more power. Consumption is brought on by this permission. Thus, it is feasible to benefit from the physical nodes' diversity by selecting the most energy-efficient servers. Figure 3 presents the overall working for the given system [28], [29].

![Figure 3. Overall working criteria of the proposed model](image)

Using the proposed algorithm section checks the physical availability of resources machine and makes distinctions between VMs based on their processor's capacity being reduced. The online machine with the smallest processor use is chosen within the procedure proceeds. This suggested algorithm employs [30], [31]. In this way, they implement and schedule VMs for data allocation which helps to reduce the energy reduction.

4. SIMULATION PROCESS

The modeling, simulation, and experimentation of novel cloud infrastructures and related application services are made possible by the flexible and extensible simulation framework known as CloudSim. Version 3.0.3 of the CloudSim simulator was utilized. The various classes that make up the CloudSim simulator serve as its building pieces. In this section, we run two data trial centers that make up the cloud computing environment in an effort to reduce energy usage while maintaining optimal operation. The way the IqrMc class functions in this regard is of relevance to us. Then, using the proposed algorithm (best energy management), we will put the VMs allocation policy (Iqr) into practice. The second step in lowering the cloud's energy usage is to put the VMs migration policy into practice. As a workload, this class always uses random policy to simulate [32], [33]. The main parameters that will be checked during the simulation process are i) The simulation's total running time equals the time it takes for all of the commands to be carried out by the VMs; ii) The energy used in this simulation is the energy used by the physical devices (hosts); iii) The number of VMs that were moved during the simulation is the total number of VMs that were
moved from one physical machine to another; iv) Performance degradation due to migration is the proportion of physical machines' performance degradation caused by the migration of VMs; v) The degree of noncompliance with the SLA contract reported at the simulation's conclusion is known as the global SLA violation; vi) The level of non-compliance with the SLA contract that was documented during the simulation is known as the average SLA violation; and vii) The quantity of actual machines changed this is the total number of physical hosts that have been shut down as a result of all VM migrations on these hosts.

5. RESULTS AND DISCUSSION

The results of multiple tests are presented in Table 1 and Figure 4. The primary metric for evaluating our method is energy utilization. From the data in Table 1 and Figure 4, it is evident that consolidation using L1 consumes less energy compared to LE consolidation. Hence, we can infer that our implementation utilizes 0.29% less energy than the previous method. Conversely, in contrast to local robust regression minimum migration time (LRRMMT), our solution demonstrates a reduction in the number of servers that need to be shut down. This is attributed to load balancing bMatching migration cost (LBBMC) having a higher number of servers with moderate loads, which is advantageous as these active servers consume less energy. Moreover, this characteristic facilitates the data center in maintaining equilibrium for an extended period.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proposed approach</th>
<th>Existent algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time of the simulation</td>
<td>903400 seconds</td>
<td>99259 seconds</td>
</tr>
<tr>
<td>Energy consumption</td>
<td>41.63 KWh</td>
<td>76.86 KWh</td>
</tr>
<tr>
<td>VMs migrated</td>
<td>42</td>
<td>5685</td>
</tr>
<tr>
<td>The decline in SLA performance due to migration</td>
<td>0.29%</td>
<td>0.33%</td>
</tr>
</tbody>
</table>

Figure 4. The result of the given proposed algorithm

Additionally, our procedure takes a long time to complete compared to the LRRMMT technique. This is a result of both the parameters of our matching technique, which is used for live migration and takes into account both the energy consumption of the VMs on the source server and the migration energy cost and the calculation time of the balancing factor F, which is not taken into account in LRRMMT. Table 2 and Figure 5, demonstrate that there are fewer migrations in LBBMC than in LRRMMT. The number of migrations raises the migration overload of VMs, which is important to minimize for the proper operation of physical machines, our reassignment technique eliminates unwanted migrations, which increases energy consumption.

A sizable portion of the migrated VMs are responsible for the performance decrease of the SLA. For the present task, these relocated VMs account for 1.23% of this degradation, compared to 4.16% in our implementation. Additionally, the total SLA violation for the existing job is 14.16%, and the overall SLA violation after applying our technique is 11.81%, indicating that our method does not adhere to all of the cloud provider's terms and conditions. Due to a significant VMs move, 1533 hosts in the current work have been shut down, which has caused a decline in SLA performance. 698 hosts are disabled as a result of our suggestion. This suggests that the migration process significantly reduces the performance of the current...
work. When compared to the current work, we find that our implementation reduces the energy consumption of the cloud for the same characteristics of a CDCs, but less closely adheres to the SLA contract's set terms. However, due to the effectiveness of its good strategy to migrate as few VMs as possible, to maintain the cloud's smooth operation, our technique offers a good alternative to repair this problem in terms of compromise between the energy consumed and the degradation of the SLA contract. The outcomes demonstrate the effectiveness of the dynamic allocation and consolidation of VMs can be used to optimize energy management in real-world CDCs, where the latter will significantly reduce greenhouse gas emissions because it minimizes energy consumption. VMs deactivate idle physical machines (hosts) and offer energy savings in the cloud computing environment. However, because disk storage and network interface play a substantial role in total energy usage, our suggested approach does not take these resources into account when reallocating VMs.

![Figure 5. The result of the given proposed algorithm](image)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proposed contribution</th>
<th>Existent algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall SLA violation</td>
<td>4.16%</td>
<td>1.23%</td>
</tr>
<tr>
<td>Average ALC violation</td>
<td>14.48%</td>
<td>11.81%</td>
</tr>
<tr>
<td>The number of physical machines switched off</td>
<td>698</td>
<td>1533</td>
</tr>
</tbody>
</table>

6. CONCLUSION AND FUTURE WORK

To reduce energy consumption, the live migration method and the blade element momentum (BEM) algorithm are integrated in this article. Performance and reducing consumption in a cloud environment. According to the simulation findings, our suggestion enhances workload management while reducing energy consumption for a significant number of servers in a data center. We plan to increase our efforts in the future to adapt our algorithm to various data centers. The next thing we want to do is investigate additional variables for improved cloud computing energy management performance and reducing consumption in a cloud environment. According to the simulation findings, our suggestion enhances workload management while reducing energy consumption for a significant number of servers in a data center. We plan to increase our efforts in the future to adapt our algorithm to various data centers. The next thing we want to do is investigate additional variables for improved cloud computing energy management.

ACKNOWLEDGEMENTS

Thanks to my SV for guidelines and support the paper financial support by the student, is a self-support paper.

REFERENCES


BIOGRAPHIES OF AUTHORS

**Soukaina Ouhame** holds her Ph.D. in Computer Science from the Faculty of Sciences of Ibn Tofail University of Kenitra. Her main research interests focus on Cloud computing, artificial intelligence, metaheuristic modeling and optimization, evolutionary computations, and optimization algorithms. She can be contacted at email: soukaina.ouhame@uit.ac.ma.

**Moulay Youssef Hadi** is a Full Professor at the Ibn Tofail University of Kenitra. He also holds the position of Deputy Director in charge of educational affairs at the Higher School of Technology of Kenitra. His research focuses on the field of Model-Driven Engineering, virtualization, and cloud computing. He can be contacted at email: hadi@uit.ac.ma.

**Amine Mrhari** earned his Ph.D. from Ibn Tofail University, Kenitra, Morocco, in 2022. Presently, he holds a position as a professor within the Moroccan Ministry of National Education, Preschool, and Sports, concurrently serving as a member of the Department of Computer Research at Ibn Tofail University, Faculty of Sciences, Kenitra, Morocco. His scholarly pursuits are primarily focused on game theory, deep learning, and resource management within the context of cloud computing. He can be contacted at email: amine.mrhari@gmail.com.

**Imane Laassar** is a Ph.D. student at the Department of Computer Science, Faculty of Science, IbnTofail University, Kenitra, Morocco. Her main research interests focus on Intrusion detection systems for internet of thing-based big data. She can be contacted at email: imane.laassar@gmail.com.