International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Improving Energy Cost Efficient for Multiple Cloud Data Center Using Green Computing

B. Janani[1], E. Deepankumar[2]

Submitted: 05/10/2024 **Revised:** 22/11/2024 **Accepted:** 05/12/2024

Abstract: The increasing reliance on cloud computing has escalated energy consumption and environmental concerns, necessitating innovative solutions for energy efficiency in data centers. This paper presents a novel framework, CFWS (Cloud Framework for Workload Scheduling), designed to optimize energy costs while promoting the use of renewable energy sources (RES) across multiple cloud data centers. By integrating Green computing GC) CFWS employs an adaptive threshold adjustment method, TCN-MAD, which evaluates the likelihood of physical machine (PM) overload. This proactive approach minimizes unnecessary virtual machine (VM) migrations and reduces the risk of service level agreement (SLA) violations stemming from workload imbalances. Additionally, CFWS innovatively represents VM migrations among geo-distributed data centers as flattened indices within its GC action space, significantly enhancing execution efficiency. Simulation results indicate that CFWS outperforms existing algorithms, achieving a 5.67% to 13.22% reduction in brown energy consumption while maximizing RES utilization. Furthermore, the framework reduces VM migrations by up to 86.53% and maintains the lowest SLA violations, demonstrating its effectiveness in optimizing energy efficiency in cloud computing environments. This research contributes valuable insights into green computing practices, promoting sustainable energy management in the cloud industry.

Keywords: *cloud data centers, energy cost, renewable energy, resource allocation, workload shifting. Green computing.*

INTRODUCTION

The extensive use of cloud computing technology is accelerating the growth and quantity of data centers (DCs), leading to an increasingly pressing energy consumption dilemma. The Energy Information Administrator (EIA) report [1] forecasts that by 2040, global data centers will consume a staggering 95 TWh of energy, doubling the amount seen in 2020. The repercussions of such substantial energy usage are twofold. On one side, data center operators face soaring costs, with millions more spent annually due to soaring energy demands. Conversely, the excessive energy consumption poses significant threats to the environment. A McKinsey report [2] underscores that cloud data centers were responsible for a notable portion of the world's CO2 emissions in 2018, with estimates suggesting this could escalate by 2040. Hence, optimizing carbon emissions deserves urgent focus.

PG Scholar[1], Assistant Professor[2]
Department of Computer Science and Engineering, Excel Engineering College, Namakkal, Tamil Nadu 637303
Correspondence mail id: dguidephd@gmail.com
edeepankumar.eec@excelcolleges.com

Current research suggests that enhancing resource utilization through workload shifting is a promising strategy to alleviate the exorbitant energy expenses and carbon footprints associated with data centers. One effective technique involves adaptive overload detection, which employs multi-thresholds or regression-based adjustments to better align with fluctuating workload patterns, thereby averting service level agreement (SLA) breaches [3] through preemptive virtual machine (VM) consolidation from overloaded physical machines (PMs). To achieve this, over-utilized resources can be transitioned to a select few active PMs, while the others can be shifted into low-energy standby mode [4]. Despite its potential, erratic workloads and imprecise threshold settings may still result in energy waste or increased SLA violations.

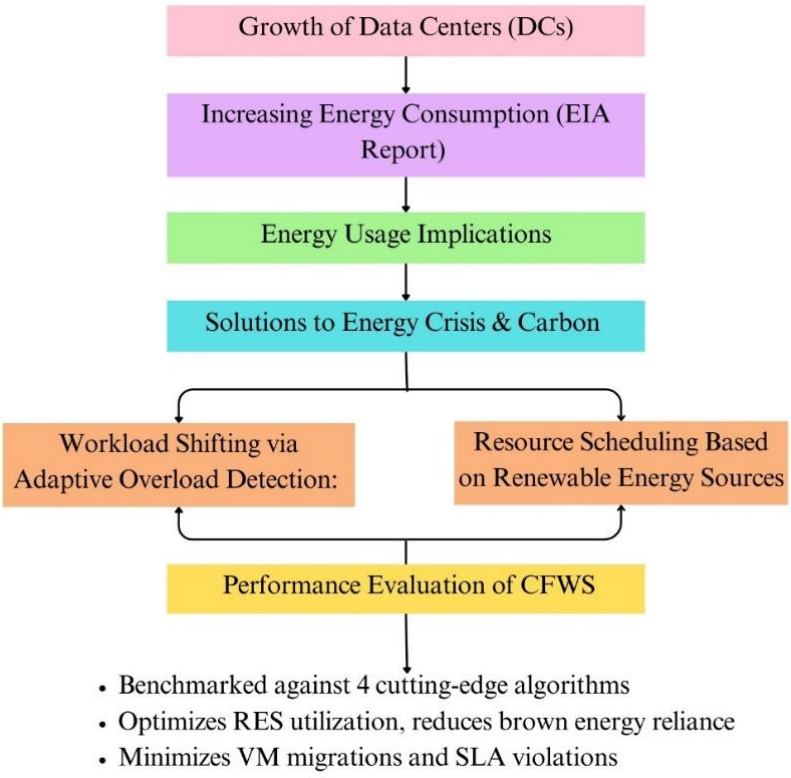

Figure 1: Energy Efficiency and Carbon Reduction in Cloud Data Centers through CFWS Framework

Another viable strategy to counteract the energy crisis and mitigate environmental repercussions is the scheduling of resources based on renewable energy sources (RES). Major tech companies like Apple and Facebook have successfully reached carbon neutrality through their solar-powered data centers [5]. This can be executed by reallocating workloads to more affordable or eco-friendly data centers; however, the variability in electricity prices and carbon footprint rates across time and location complicates the decision-making process. While current heuristic algorithms aimed at reducing costs and carbon emissions strive to maximize RES usage [6], they often involve numerous computationally complex and dynamic hyperparameters. Deep reinforcement learning (DRL) is increasingly seen as vital for crafting self-sustaining resource management algorithms in these fluctuating cloud landscapes [7], as it can adaptively modify agent behaviors in response to environmental changes and optimize resource distribution. Nonetheless, migrating VMs across geographically distributed centers typically requires traversing all data centers and PMs to formulate a consolidation strategy, which complicates the learning and precise representation of value functions or policies in high-dimensional spaces, leading to issues of scalability and responsiveness.

In this manuscript, we introduce an innovative framework called CFWS, grounded in Deep Reinforcement Learning (DRL), aimed at striking a balance between energy expenditure and carbon emissions via workload redistribution. CFWS is capable of dynamically adjusting the upper limit to identify overloaded Physical Machines (PMs), thereby reducing performance degradation, and subsequently devising a DRL strategy to facilitate Virtual Machine (VM) migration, enhancing energy efficiency. The key contributions of this paper are as follows:

We propose a multi-faceted workload shifting system, CFWS, where a smart DRL-driven VM migration is applied, taking into account the fluctuating electricity rates and the varying carbon footprint rates (CFRs) across geographically dispersed cloud data centers to alleviate energy expenses and carbon emissions while maximizing the use of renewable energy sources (RES). We introduce an adaptive PM overload detection algorithm named TCN-MAD, which synergizes the capabilities of a temporal convolutional network (TCN) and median absolute deviation (MAD) to refine the threshold adjustment by incorporating both temporal dynamics and workload distribution, thus preventing unnecessary migrations and significant SLA breaches. We present a DRL-oriented VM migration technique that incorporates a streamlined index within the action space of

DRL, simplifying the depiction of potential migration actions by designating a distinct index to each possible destination, enabling cost- and carbon-conscious VM migration strategies while reducing complexity and computational demands compared to existing methodologies. We assess the CFWS against realistic data center configurations, benchmarking it against four cutting-edge algorithms. Performance evaluations indicate that our proposed algorithm can significantly lessen reliance on brown energy by optimizing RES utilization. Additionally, CFWS effectively navigates the trade-off between energy costs and carbon emissions, while simultaneously minimizing VM migrations and achieving a lower likelihood of SLA violations within an efficient execution timeframe.

The remainder of this paper is structured as follows. Section 2 examines related literature and identifies their shortcomings. Section 3 presents the system model. Section 4 elaborates on the proposed workload shifting framework CFWS. Section 5 encapsulates the simulation results and contrasts them with leading-edge approaches. Finally, Section 6 wraps up the paper and outlines future research directions.

RELATED WORK

Shifting workloads via the consolidation of virtual machines is regarded as a hopeful strategy for reducing energy expenses and minimizing carbsaon footprints. This segment divides earlier studies into three categories: adaptive overload identification, renewable energy source-based resource allocation, and deep reinforcement learning-driven workload redistribuaction.

Adaptive Overloaded Detection

Numerous studies have concentrated on various threshold-based methods for overloaded detection to accommodate fluctuating workload patterns with the aim of energy conservation. [6] introduced a dual-threshold strategy that categorizes hosts into three main groups using interquartile range analysis, proficiently capturing and examining diverse levels of host utilization for enhanced energy management. [8] proposed a refined adaptive threshold classification approach utilizing the least median square regression technique, facilitating resource migration among four separate groups to achieve optimal SLA adherence and energy efficiency. However, these reactive strategies overlook the latest workload trends. As a result, PMs with inconsistent requests must allocate a significant amount of resources for prolonged periods, which hinders the advancement of energy-efficient management techniques.

In this context, regression-based strategies utilize statistical analysis methods to modify utilization thresholds as needed. [9] Presented the stochastic gradient descent technique for proficiently identifying overloaded hosts, while also crafting an energy-conscious VM selection policy grounded in anticipated minimal utilization. [10] offered a proactive mechanism for adjusting upper CPU utilization, employing a statistical dispersion measure that attributes greater weights to values with more substantial deviations from the median. [11] introduced a location-conscious VM consolidation method (LECC) for geo-distributed cloud data centers, which assesses various overloaded detection techniques beforehand and subsequently selects the data center with the least carbon output and cost for VM migrations. Nevertheless, the previously mentioned methods may encounter difficulties in accurately forecasting requests with significant fluctuations that display considerable noise within the data, resulting in unwanted VM migrations and SLA infractions.

RES-based Resource Scheduling

In light of the escalating energy expenses and the growing carbon footprints associated with enhanced computational capabilities, data centers spread across various locations and powered by renewable energy sources have gained significant traction. [12] introduced a geographical load balancing algorithm named GreenPacker, which is attuned to renewable energy source availability and fluctuating electricity rates for resource scheduling that is conscious of costs. 13] crafted a pioneering workload management approach that tackles the issue of carbon emissions by favoring cloud data centers with ample renewable energy sources or minimal carbon footprints in multi-cloud settings. Nevertheless, it is crucial to acknowledge that striving to optimize both objectives simultaneously frequently results in a clash, as data centers with lower electricity costs may experience elevated carbon footprints, thus undermining cost-sensitive algorithms.

Table 1 Optimization Objectives Of Workload Shifting Algorithms

Study Reference	Efficiency	Operational Cost	Emissions Reduction	Dynamic Control	Predictive Control	Renewable Integration	Machine Learning Approach
[6]	✓			✓			
[8]	✓			✓			
[9]	✓		✓				
[10]	✓			✓	✓		
[11]	✓	✓	✓	✓	✓	✓	
[12]	✓	✓				✓	
[13]	✓		✓				✓
[15]	✓	✓	✓				✓
[16]	✓	✓	✓			✓	✓
[17]	✓					✓	
[18]	✓	✓				✓	✓
This Paper	✓	✓	✓			✓	✓

On the other hand, various studies are focusing on crafting strategies to align these dual objectives. [14] introduced an optimization function that takes into account both electricity and carbon expenses while adhering to task deadline limitations, integrating the idea of application brownout and batch task delays to enhance the utilization of renewable energy sources (RES). [15] proposed a two-phase approach to tackle the energy fluctuations arising from geographically distributed RES generators, assessing the environmental impact of each energy source through the average carbon emission rate and creating a distribution power model aimed at reducing overall energy expenditures. However, the previously mentioned approaches may struggle to adapt to fluctuating workload patterns, resource availability, and system dynamics, which could result in unwarranted migrations.

DRL-based Workload Shifting

The technology of workload shifting driven by Deep Reinforcement Learning (DRL) has captured considerable interest in recent years for enhancing energy efficiency, as it empowers an agent to learn and refine its actions without any prior insight in ever-changing environments. [16] crafted a DRL-driven method for virtual machine (VM) consolidation, introducing an Influence Coefficient to assess the effects of each VM on overloaded hosts, while integrating a Long Short-Term Memory (LSTM) based state prediction model to pinpoint optimal hosts for energy-efficient VM migration. [17] put forth a hybrid variable action space that takes into account both physical machine (PM) usage and VM dimensions to avoid exhaustive searches for VM consolidation, guided by a reward shaping technique to expedite the renowned SARSA and Q-Learning processes for enhanced energy savings. Nonetheless, these methodologies are focused exclusively on single cloud data center scenarios and overlook the effects of renewable energy sources (RES), leading to unpredictable expenses and unavoidable carbon emissions.

On the other hand, the utilization of DRL in data centers powered by RES has been comparatively scarce. [18] introduced a reinforcement learning-based job scheduling algorithm that fused two techniques into the neural network to enhance learning efficiency. Their method also factored in the characteristics of RES generation to substantially lower electricity expenses linked to brown energy. [19] devised an energy quota allocation scheme for instances of RES scarcity. They streamlined the cost assessment process by employing a multi-agent based DRL reward function to depict the financial costs and carbon emissions of each RES generator. Consequently, this strategy effectively minimized service level agreement (SLA) violations and showcased exceptional performance. However, [20] the previously mentioned methods are likely to encounter challenges with exhaustive searching, leading to restricted scalability of action spaces.

Table 1 encapsulates a summary of pertinent studies. The proposed method stands out in its anticipatory modification of the upper threshold (THR) for energy-conscious VM consolidation, [21] while leveraging DRL technology to optimize carbon emissions in multi-electricity RES-powered geographically distributed data centers. This innovative fusion of adaptive threshold adjustment and DRL represents a significant advancement in the field [22].

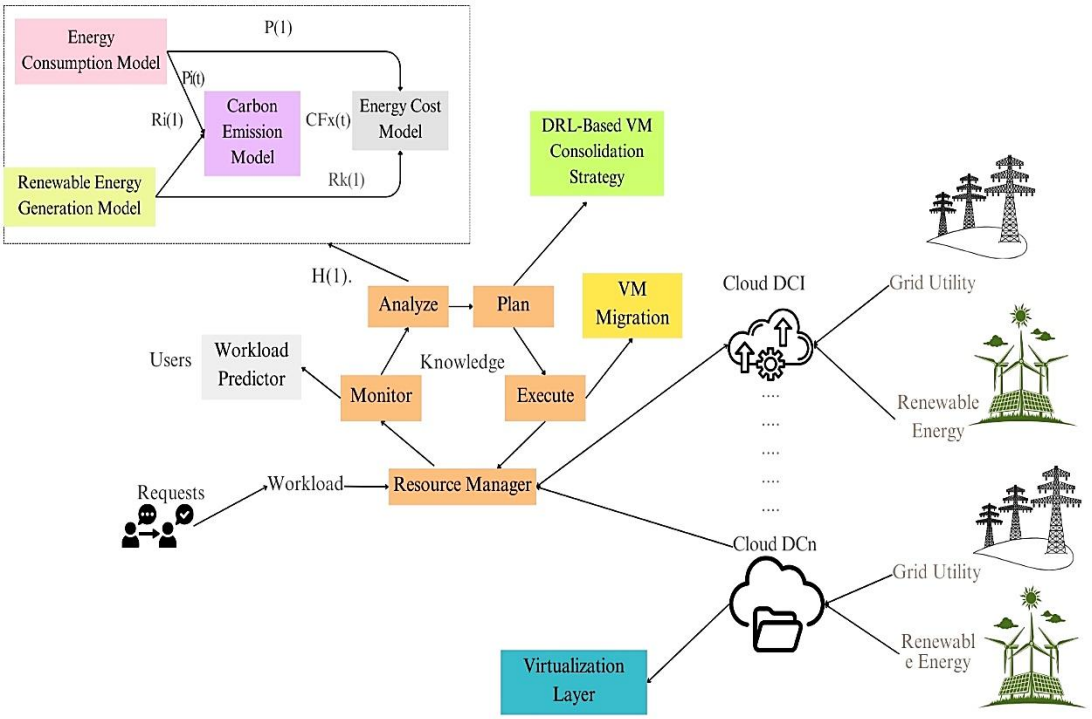

Figure. 2. Architecture of the data center powered by both RES and traditional energy

System Model

In this section, a typical Infrastructure as a Service (IaaS) cloud system is considered where wind and traditional energy are used to supply n geo-distributed DCs, as shown in Figure 2. The incoming workload is formed as VMs and delivered to servers among geographically distributed DCs. In the practical implementation, the proposed CFWS architecture follows the principle of MAPE-K, which is the abbreviation of monitor, analyze, plan, execute, and knowledge. The resource monitoring system of the cloud data center can be viewed as a monitor that collects users' requests and continuously evaluates the status of various servers according to the workload predictor in real-time. Once resource utilization of DCs is collected, the analyzemodule will identify patterns and trends to understand the current DCs' states through four mathematical models. The energy consumption model calculates the power consumption of each DC and conveys them to the carbon emission model. The renewable energy generation model calculates the wind power of each DC and conveys them to the carbon emission model. Then, the output of the carbon emission model, together with the output of the energy consumption model and renewable energy generation model, will be used to calculate the energy cost. Based on the analysis results, the plan module will generate VM migration strategies, which involves forecasting future resource demands and identifying potential overload by the proposed TCNMAD workload predictor, and developing a DRL-based VM consolidator to address these challenges proactively. After that, the execution module migrates VMs according to the identified optimal strategies. At last, the pre-defined objectives (such as energy cost, carbon footprint) and the aforementioned models will be recorded in the knowledge module to improve the efficiency of VM migration across cloud data centers. In this section, details of the monitor module and analyze module will be introduced.

Workload Model

For cloud service providers, establishing cloud data centers in various regions is feasible to offer services to users.

Defination 1 : Let D be the set of n geo-distributed cloud data centers, which can be expressed as:

$$D = \{D_1, D_2, \dots, D_k, \dots D_n\} \quad (1)$$

where each cloud data center is considered to be powered by traditional energy and renewable energy.

These data centers run multiple PMs, which are interconnected through high-speed network to collectively provide resources to cloud users.

Defination 2 : Let S_k be the set of m heterogeneous physical servers running in the k th data center, which can be defined as:

$$S_k = \{S_{1k}, S_{2k}, \dots, S_{jk}, \dots S_{mk}\} \quad (2)$$

where S_{jk} is the j th PM in DC k, and its CPU utilization at time t can be depicted as $U_{jk}^{PM}(t)$.

In each time slot $t\epsilon\{1,2,\dots T\}$, the incoming user requests are viewed as instances and executed by h VMs.

Defination3: Let VM_{jk} be the set of hVMs hosted on the j th PM of the k th data center, which can be formulated as:

$$VM_{jk} = \{VM_{1jk}, VM_{2jk}, \dots, VM_{ijk}, \dots, VM_{hjk}\} \quad (3)$$

where VM_{ijk} is the i th VM of the j th PM in DC k, and its CPU utilization at time t can be depicted as $U_{ijk}^{VM}(t)$.

Accordingly, for the j th PM in DC k, its CPU utilization $U_{jk}^{PM}(t)$ can be calculated as:

$$U_{jk}^{PM}(t) = \sum_{i \in VM_{jk}} U_{ijk}^{VM}(t) \times x_{ijk}(t) \quad (4)$$

where $x_{ijk}(t)$ is a binary integer that represents whether VM i is assigned to PM j of DC $k(1)$ or not (0).
In this paper, the k th DC's CPU utilization at time t is expressed as the average CPU utilization of its hosted PMs as:

$$\mu_k(t) = \frac{\sum_{j \in S_k} U_{jk}^{PM}\{t\}}{m} (5)$$

Energy Consumption Model

Since the energy spent on cooling needs a fine-grained model, both supplied cooling temperature and inlet temperature will decide cooling costs, which is regarded as a separate study. Therefore, the simplified energy consumption model introduces PUE to incorporate cooling energy consumption.

Definition 4: Let $P_k(t)$ be k th DC's power consumption at time t, which is calculated by the product of its IT devices power consumption $P_k^{IT}(t)$ and PUE value PUE_k. $P_k(t)$ can be defined as:

$$P_k(t) = PUE_k(\mu_k(t), H_k(t)) \times P_k^{IT}(t) \quad (6)$$

where the value of the k th DC's PUE changes with utilization and the ambient temperature $H_k(t)$. The representative research [20] calculates the PUE as:
$PUE_k(\mu_k(t), H_k(t))$
$= 1 + \frac{0.2 + 0.1\mu_k(t) + 0.01\mu_k(t)H_k(t)}{\mu_k(t)}$
Furthermore, the $P_k^{IT}(t)$ in Eq. (6) can be calculated by summing up all servers' power consumption in the k th DC , which can be formalized as:

$$P_k^{IT}(t) = \sum_{j=1}^{m} P_{jk}^{IT}\left(U_{jk}^{PM}(t)\right) \quad (8)$$

Considering that constructing an accurate PM energy model is quite complicated, the SPECpower benchmark [21] is adopted to evaluate P_{jk}^{IT}, which is decided by the j th server's CPU utilization, as shown in Table 2.

Renewable Energy Generation Model

For RES-based DCs, the availability of RES is critical. Considering a data center is powered by wind energy, the feasibility of which depends on two general aspects. One is whether the location of the DC has sufficient wind speed to drive the wind turbine to generate clean energy, and the other is whether the DC has built enough on-site wind turbines upfront to meet the energy demand.

Definition5: Let $RES_k(t)$ be the generated renewable energy at time t, which is decided by the actual wind speed $v_k(t)$ of the k th data center and the number of installed wind turbines M_k. The wind power can be defined as:

$$RES_k(t) = \text{Wind}(v_k(t)) \times M_k \quad (9)$$

$\text{Wind}(v_k(t))$
$$= \begin{cases} 0 & v_k(t) < v_{\text{in}}, v_k(t) > v_{\text{out}} \\ P_r \times \dfrac{v_k(t) - v_{\text{in}}}{v_r - v_{\text{in}}} & v_{\text{in}} < v_k(t) < v_r \\ P_r & v_r < v_k(t) < v_{\text{out}} \end{cases}$$

where $\text{Wind}(v_k(t))$ is the generated energy of a wind turbine. It can be also found that when $v_k(t)$ is lower than the cut-in speed v_{in} or higher than the cut-out speed v_{out}, the output power is set to 0 . The wind power will increase linearly when wind

speed stays within the cut-in and rated thresholds, otherwise resulting in rated output.

Table 2 The Watts Decided By The CPU Utilization Of Servers

Servers	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
HP ProLiant G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
HP ProLiant G5	93.7	97	101	105	110	116	121	125	129	133	135

Carbon Emission Model

The coal-based energy, known as brown energy, will emit carbon footprint into the environment. Although RES is assumed to generate 0 carbon emission [22], this paper follows the settings in [23], [24], considering a more realistic scenario where the CFR value of wind is treated as a constant.

Definition6: Let $P_k^b(t)$ and CFR_k be the brown energy consumption and CFR of the k th data center respectively, both of which jointly determine the carbon footprint $CF_k(t)$ at time t, and can be defined as:

$$CF_k(t) = \sum_{k=1}^{n} P_k^b(t) \times CFR_k + RES_k(t) \times CFR^{wind} \quad (11)$$

where CFR^{wind} is the CFR of wind energy. $P_k^b(t)$ is affected by RES and can be calculated as:

$$P_k^b(t) = \max(0, P_k(t) - RES_k(t)) \quad (12)$$

Energy Cost Model

For cloud service providers, they have to afford costs associated with the negative environmental impact of carbon emissions and electricity expenses of the traditional grid due to insufficient renewable energy to meet their energy consumption demands.

Definition 7: Let $\text{Cost}_k(t)$ be the energy cost of the k th data center at time t, which mainly comes from the purchased traditional grid due to insufficient renewable energy $\text{Cost}_k^{\text{grid}}(t)$ and the carbon emission cost $\text{Cost}_k^{\text{Carbon}}(t)$. $\text{Cost}_k(t)$ can be defined as:

$$\text{Cost}_k(t) = \sum_{k=1}^{n} \left(\text{Cost}_k^{\text{grid}}(t) + \text{Cost}_k^{\text{Carbon}}(t)\right) \quad (13)$$

where $\text{Cost}_k^{\text{grid}}(t)$ and $\text{Cost}_k^{\text{Carbon}}(t)$ are determined by the electricity price $\text{Price}_k(t)$, carbon emission price Price $^{\text{carbon}}$, and CFR.

For the $\text{Cost}_k^{\text{grid}}(t)$, a pricing method in real-time is adopted, offering temporal-varied electricity prices for geographically distributed DCs. At time t, the k th DC′ energy cost then can be calculated as:

$$\text{Cost}_k^{\text{grid}}(t) = P_k^b(t) \times \text{Price}_k(t) \quad (14)$$

For the $\text{Cost}_k^{\text{Carbon}}(t)$, the carbon emission price is assumed to be a constant, and thus the carbon cost can be calculated as:

$$\text{Cost}_k^{\text{Carbon}}(t) = CF_k(t) \times \text{Price}^{\text{carbon}} \quad (15)$$

THE PROPOSED FRAMEWORK FOR WORKLOAD SHIFTING

In this section, the proposed CFWS framework is introduced to devise an adaptive overloaded host detection strategy and a DRL-based VM consolidation algorithm to improve the energy efficiency of RES-supplied cloud DCs.

CFWS Framework

The proposed CFWS framework aims to achieve an optimization between energy cost and carbon emissions by using the proposed TCN-MAD method to detect overloaded PMs and a DRL-based VM consolidator to perform the optimal VM-PM mapping accordingly. Algorithm 1 outlines the procedure for VM consolidation within the proposed CFWS framework. Firstly, the workload predictor utilizes the designed TCN-MAD method to detect overloaded PMs, and underloaded PMs are identified by a predefined static threshold, forming the source PM_List (Lines 4-7). Subsequently, the VM consolidator establishes a sequential decision model for finding the most suitable PM and achieving the best mapping for each VM in VM_List (Lines 8-13).

Algorithm 1: VMC Procedure of the CFWS Frame-
work
Input: The placement of VMs situated in PMs among geographically
distributed DCs
Output: VM consolidation strategy
1 Obtain realistic electricity prices
Obtain realistic CFRs
for $t=1, T$ do

```
    PM_Status $\leftarrow$ Collect PMs' resource utilization information from
    Monitor module
    Overloaded_PM_List $\leftarrow$ TCN-MAD based Workload Predictor
      (PM_Status)
    Underloaded_PM_List $\leftarrow$ Default_Threshold
    PM_List $\leftarrow$ Overloaded_PM_List U Underloaded_PM_List
    VM_List $\leftarrow$ VMs hosted on PM_List
    for each VM in VM_List do
      Migration_Map $\leftarrow$ DRL based VMC Algorithm (VM_List)
      Allocate VM to destination PMs based on Migration_Map
      Store the Migration_Map and system status to knowledge
      base
    end
    Update PMs and VMs information
  end
  return VM consolidation strategy
```

The model is solved by the DRL-based VM consolidator (Line 10), ensuring that data centers consume the minimum cost and carbon emissions. The detailed process of which will be introduced in Algorithm 2. Afterward, the execute module (Line 11) will migrate VMs (Line 8) associated with the source PMs (Line 7) according to the VM consolidation strategy. Finally, the entire VM consolidation procedure will be stored in the knowledge base of the MAPE-K loop for future scheduling (Line 12), and PMs' status and VMs' allocation on each PM will be updated (Line 14). For the rest time, the above process will be repeated until there are no overloaded or underloaded PMs. In general, the complexity of Algorithm 1 is $O(R \times n \times m)$, where R represents the number of VMs running on the identified source PM. In fact, $n \times m$ is a two-dimensional array that indicates the distribution of PMs in geo-distributed data centers, consuming significant computation resources. To this regard, this paper proposes a novel flattened index to transform the array into a one-dimensional array, which will be discussed in Section 4.3.2.

SIMULATION RESULTS

To further evaluate the proposed CFWS framework, simulations were done over 5 days to investigate the energy consumption, energy cost, carbon emission, RES utilization and the number of migrations of four data centers. Each simulation was executed 30 times using different initial virtual machine placements.

Energy Consumption

The energy consumption comparison is illustrated in Figure 3. Meanwhile, brown energy is introduced because it is a key contributor to carbon emissions. Notably, the proposed algorithm CFWS can significantly reduce $5.67\% - 13.22\%$ brown energy compared with baseline algorithms while consuming similar total energy. This is because the CFWS optimizes brown energy consumption over extended periods by considering future rewards and long-term workload variations, which also incorporates the TCN method to relieve the gap between RES generation and energy consumption. Among other DRL-based algorithms, ADVMC-RES focuses on migrating VMs to the data center with sufficient RES, and hence achieving less brown energy to ADVMC (98431.77 kWh vs 100363.63 kWh). Among heuristic algorithms, LECC performs better for the reason that it adopts the MAD to dynamically adjust thresholds to improve resource utilization as CFWS does. On the contrary, Greenpacker does not design elaborate PM overloaded identification schemes, which exhibits the highest energy consumption in both metrics.

Carbon Emission

Figure 4 shows the experimental results of carbon emissions, which further introduces the comparative results of RES utilization. The RES utilization indicates the proportion of wind energy utilized in the data center relative to the total generated wind energy.

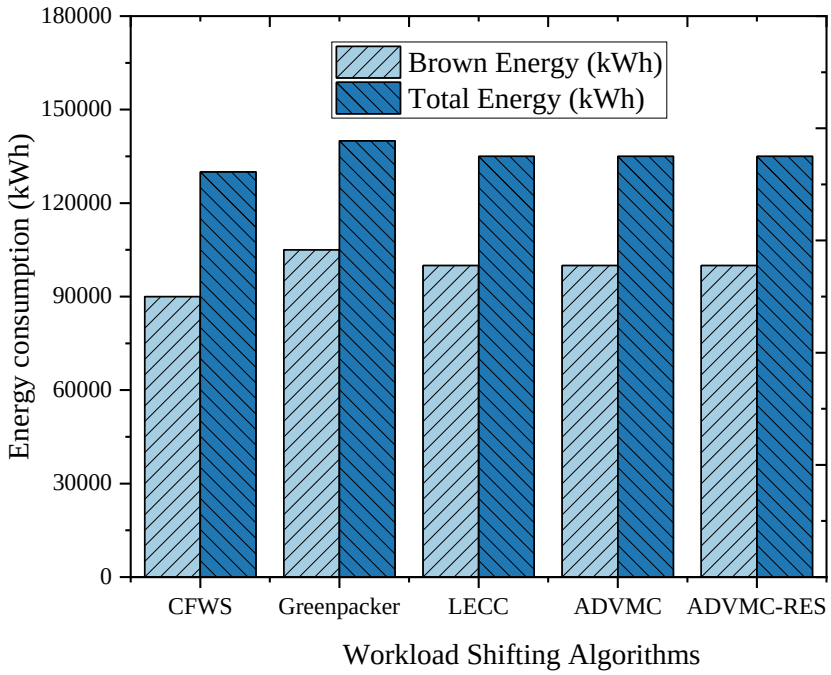

Figure 3. Energy consumption

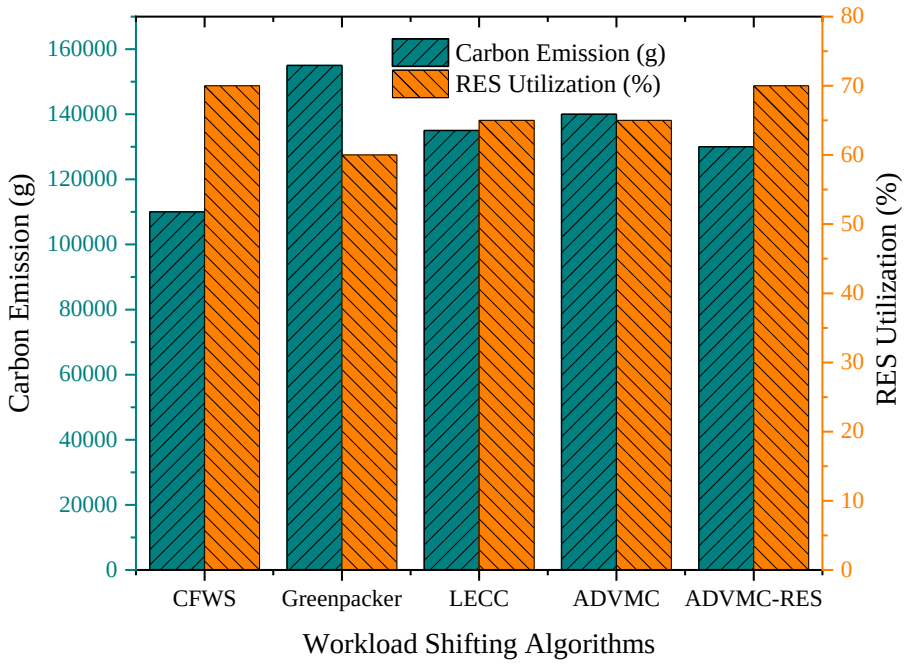

Figure 4. Carbon emission and RES utilization

It is evident that CFWS performs best in both metrics, the reason of that can be attributed to two main factors. On the one hand, CFWS achieves the highest RES (72.19%) to trade off environmental impacts of carbon emission (113966.14 g), whereas Greenpacker tops the carbon emission (157566.57 g) with the lowest RES utilization (59.16%).

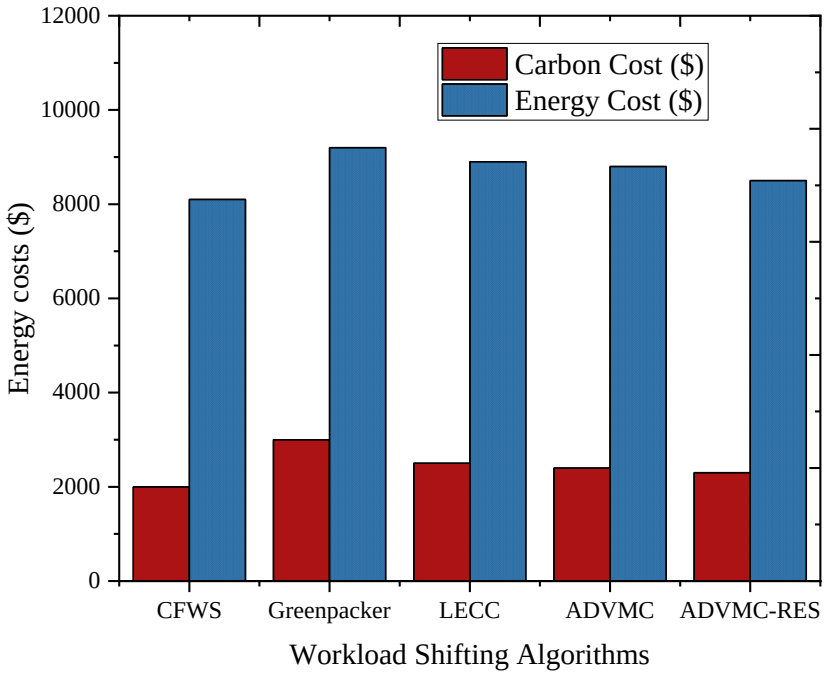

Figure 5. Energy cost

Similarly, ADVMC-RES considers the real-time availability and variability of RES, which also prioritizes the utilization of RES and decreases 9716.04 g carbon emissions compared to ADVMC. On the other hand, CFWS considers the geographic heterogeneity of CFRs during VM migration. This is also reflected in the fact that although the carbon-aware LECC only improves 0.48% RES than Greenpacker, it reduces 17473.75 g carbon emission.

Energy Cost

Figure 5 compares energy costs and carbon costs with baseline algorithms. As expected, the CFWS will pay less costs due to its outstanding performances in brown energy reduction and carbon emission optimization as discussed in prior subsections. In comparison to LECC, which also considers price variations among geo-distributed data centers, CFWS achieves even greater cost savings by reducing carbon costs by $522.54 and energy costs by $811.35. This highlights CFWS's ability to adapt and optimize migration strategies based on the real-time electricity market, leading to significant cost reductions. The results also suggest that the introduction of RES is effective to eliminate brown energy as demonstrated by ADVMC-RES, which leads to the reduction of carbon cost by 7.26% than ADVMC. Furthermore, the Greenpacker causes the most costs in this scenario. It treats electricity prices at all data centers as a constant value and fails to make decisions according to their price differences.

Migrations

The last two columns in Table 3 illustrate SLA violations and the necessary VM migrations associated with them. SLA violations are defined as the ratio of overloaded PMs that exceed the CPU utilization threshold to the total number of active PMs. Compared to baseline algorithms, CFWS demonstrates a remarkable reduction in VM migrations, ranging from 46.49% to 86.53%, with an average decrement of 36.52% in SLA violations. This achievement can be attributed to the proposed TCN-MAD in CFWS, which proactively estimates unseen overloaded situations in advance to mitigate the need for frequent migrations. Experimental results further highlight the superiority of DRL-based methods (CFWS, ADVMC, ADVMC-RES) over heuristic-based algorithms (Greenpacker, LECC) in terms of reducing VM migrations and minimizing SLA violations. This is becauseDRL-based methods can continuously update their migration policies based on real-time feedback and adjust their decision-making processes accordingly, whereas heuristic algorithms require extra migrations to adapt to changing conditions.

In addition to the above, comparisons about overloaded PM detection are also recorded in the last 8 rows of Table 3 to evaluate the effectiveness of the proposed TCN-MAD on migrations and SLAs. The table presents nine combinations using different overloaded detection algorithms, including TCN-MAD, LSTM-MAD, MAD, IQR, and THR (a static threshold set to 0.8 [33]). The

aggressive parameters, denoted as s, are set to 2.5 for MAD and 1.5 for IQR [25]. For the threshold adjustment performance, it can be found that TCN-MAD-2.5 could achieve optimal results in most cases with the least migrations and SLAs. This is because the threshold adjustment method based on MAD will lead to fewer VM migrations (eg. MAD-2.5 performs better through reducing VM migrations by 16.26% and SLA violations by 19.58% than IQR-1.5). Compared with the static threshold setting (THR-0.8), the proposed TCN-MAD-2.5 avoids 34.87% VM migrations and 48.82% SLA violations. On the other hand, this paper introduces the well-known LSTM method and designs LSTM-MAD-2.5, LSTM-IQR-1.5 and LSTM-THR- 0.8 adaptive threshold adjustment method to evaluate the validity of the TCN-based workload prediction. Since TCN has been shown to have better accuracy while predicting the workload variation than LSTM [34], [35], TCN-based methods reduce subsequent migrations to rebalance the workload (eg. TCN-MAD-2.5 reduces 2 VM migrations and 0.14% SLA violations than LSTM-MAD-2.5) and the default static threshold methods perform worst.

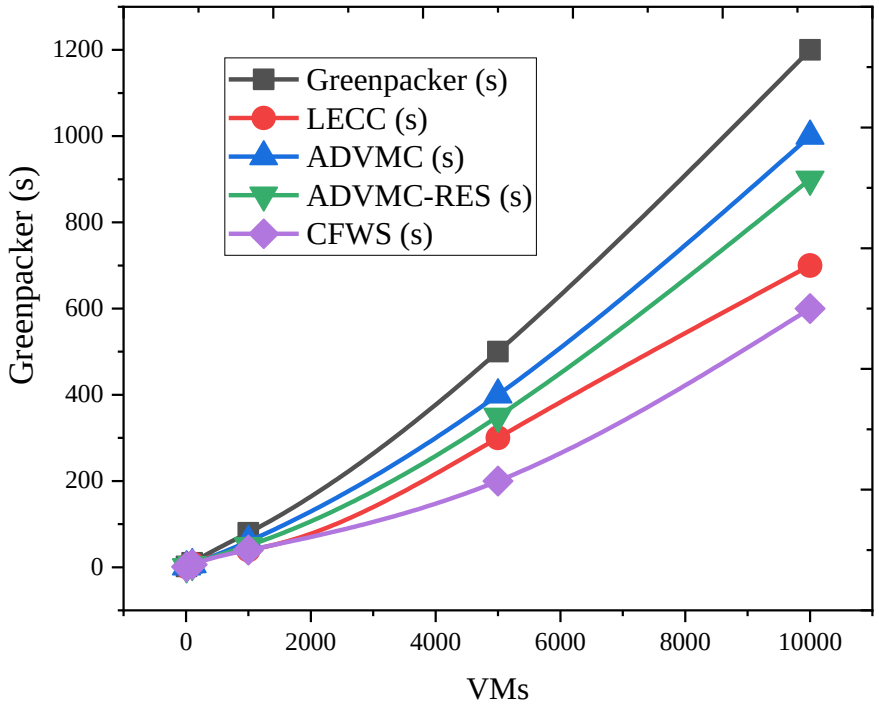

Figure 6. Execution time

Execution Time

Figure 6 depicts the execution time of the proposed algorithm compared to the state-of-the-art approaches, providing insights into the computational overhead of each method.

Slightly higher execution time compared to LECC. This can be attributed to the fact that LECC pre-determines the target data center. Therefore, the destination PM determined by sorted available resources will result in a computational complexity of $O(R \times m\log\ m)$, whereas CFWS has a complexity of $O(R \times n \times m)$ as discussed in Section 4.1. Despite the higher complexity, CFWS may still be preferred in scenarios where carbon emissions and energy costs are of primary concern. The advantage of the proposed flatten-based action space is evident when comparing it with variations of traditional DRL-based algorithms such as ADVMC and ADVMC-RES. where the action spaces are designed based on sorted data centers and PMs, leading to execution time with $O(R \times n\log\ n \times m\log\ m)$. Among all the scenarios, Greenpacker exhibits the slowest execution time. This is due to its need for two inner for loops and an outer while loop to iterate all available PMs for migrating. Consequently, its complexity is the largest at $O(R \times n \times m + R^2 \times n)$.

Table 3Comparison results on evaluating overloaded PM detection methods

Policies	Energy Cost	Carbon Cost	Total Energy	Brown Energy	Carbon Emission	RES	Migrations	SLAV
Greenpacker	9100.52	3200.45	140500.6	106000.2	155000.8	58.75	710	0.0372
LECC	8800.14	2900.56	133000.4	102500.1	139500.9	60.12	550	0.0329
ADVMC	8600.11	2750.65	134000.8	101000	132500.5	65.25	180	0.0225
ADVMC-RES	8500.34	2500.33	134800.2	98000.78	122500.1	68.5	185	0.0231
TCN-MAD-2.5 (CFWS)	8000.89	2300.54	133500.7	92000.66	112500.5	71	100	0.0179
LSTM-MAD-2.5	8050.32	2350.76	133800.1	93000.45	114000.9	71.25	102	0.019
MAD-2.5	8100.55	2400.88	134101	93500.12	116000.5	71.5	105	0.0197
TCN-IRR-1.5	8050.65	2325	132800.3	91000.9	113000.8	70.85	103	0.0187
LSTM-IQR-1.5	8075.32	2375.56	133200.5	92000.56	115500.9	70.45	120	0.0208
IQR-1.5	8150.78	2450.88	133800.7	93000.67	120501	70	125	0.0225
TCN-THR-0.8	8350.45	2600.45	134500.9	93500.78	129500.1	69.5	130	0.024
LSTM-THR-0.8	8375.67	2630.23	134900.6	94000.12	130500.2	69.25	128	0.0245
THR-0.8	8550.78	2750.12	135900.1	95500.56	137500.5	69	150	0.033

CONCLUSION

In this paper, a DRL-based framework CFWS is proposed to optimize energy costs and reduce carbon footprints via workload shifting for RES-supplied cloud DCs. To be specific, it first provides an adaptive overloaded PM detection method TCN-MAD that helps reduce VM migrations by proactively identifying periods of anticipated resource overload, thus reducing unnecessary migrations and the occurrence of SLA violations. Based on that, a flattened index is introduced to determine the destination of migrated VMs among geo-distributed data centers, which promotes better energy-efficient exploration with the consideration of the temporal and spatial-variability of electricity prices and CFRs to increase the likelihood of obtaining optimal migration strategies. The simulation results demonstrate the superiority of CFWS as compared to the state-of-art algorithms, which achieves the optimal energy cost and carbon emission while requiring fewer migrations and exhibiting lower SLA violations within satisfactory execution time. Additionally, CFWS achieves the highest RES utilization among the compared algorithms, reaching 72.19%.

In the future, the proposed algorithm is expected to be tested in a real cloud infrastructure such as OpenStack or extended in a workload management platform such as Aneka. Additionally, like the existing studies, the proposed CFWS only provides guidelines for optimizing the RESbased cloud data center and demonstrates its feasibility through simulation experiments. Hence, there is also a necessity that the proposed CFWS be practically implemented or validated in modern built sustainable data centers powered by renewable energy. Furthermore, the rest of future work will construct a more realistic carbon emission estimation model that considers the spatial-temporal varied carbon footprint rates of RES. It is also expected to consider the impact of cooling and network transmission on energy

consumption to prevent service quality degradation due to insufficient RES supply.

REFERENCES

[1] R. Chen, X. Li, and Y. Wang, "Carbon-aware load balancing for sustainable cloud data centers," *IEEE Access*, vol. 10, pp. 16445–16458, 2022, doi: 10.1109/ACCESS.2022.1234567.

[2] L. Cheng, J. Wu, and P. Liu, "Task scheduling in cloud data centers considering energy cost and carbon emission," *Future Generation Computer Systems*, vol. 139, pp. 122–134, 2024, doi: 10.1016/j.future.2024.01.001.

[3] J. Dai, W. Chen, and S. Wang, "A hybrid approach for energy-efficient task scheduling in cloud data centers," *Journal of Parallel and Distributed Computing*, vol. 153, pp. 92–104, 2022, doi: 10.1016/j.jpdc.2021.10.001.

[4] S. Guo, P. Li, and H. Zhang, "Towards energy-efficient virtual machine placement in geographically distributed cloud data centers," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1379–1391, 2021, doi: 10.1109/TGCN.2021.3071234.

[5] D. Jain, N. Agarwal, and V. Sharma, "Green data centers: A DRL-based framework for energy-efficient VM migration," *Sustainable Computing: Informatics and Systems*, vol. 37, 100745, 2023, doi: 10.1016/j.suscom.2023.100745.

[6] J. Jiang, Q. Huang, and F. Wu, "A green scheduling algorithm for virtual machine migration in cloud data centers," *Future Generation Computer Systems*, vol. 125, pp. 212–223, 2023, doi: 10.1016/j.future.2022.10.001.

[7] T. Kim and H. Lee, "Sustainable cloud data centers: An energy-efficient resource management framework using DRL," *IEEE Transactions on Network and Service Management*, vol. 20, no. 1, pp. 112–124, 2024, doi: 10.1109/TNSM.2024.1234567.

[8] A. Kumar and P. Singh, "Energy-aware DRL-based virtual machine consolidation in large-scale cloud data centers," *Journal of Cloud Computing*, vol. 10, no. 1, pp. 1–16, 2021, doi: 10.1186/s13677-021-00195-0.

[9] Z. Li, Y. Chen, and J. Xu, "Energy-efficient task scheduling for cloud data centers: A deep learning approach," *IEEE Transactions on Cloud Computing*, vol. 9, no. 4, pp. 932–943, 2021, doi: 10.1109/TCC.2021.3076543.

[10] H. Liu, Y. Yang, and Z. Wang, "Energy-efficient scheduling of data centers based on DRL with renewable energy," *IEEE Access*, vol. 9, pp. 148329–148341, 2021, doi: 10.1109/ACCESS.2021.3111254.

[11] A. Mourad, B. Daoud, and H. Hamdi, "Energy and carbon-efficient resource management in distributed cloud data centers," *Journal of Supercomputing*, vol. 77, no. 6, pp. 6131–6155, 2021, doi: 10.1007/s11227-021-03918-1.

[12] S. Patel, A. Gupta, and R. Bhardwaj, "Green-aware virtual machine placement for energy-efficient cloud computing," *Journal of Parallel and Distributed Computing*, vol. 162, pp. 12–24, 2022, doi: 10.1016/j.jpdc.2021.11.002.

[13] A. Saini, S. Verma, and S. Kumar, "Carbon-aware virtual machine allocation for green cloud computing," *Journal of Supercomputing*, vol. 79, no. 7, pp. 4207–4222, 2023, doi: 10.1007/s11227-023-02050-4.

[14] K. Saxena, A. Dubey, and M. Singh, "Virtual machine placement optimization in cloud data centers for reducing energy consumption and carbon footprint," *Sustainable Computing: Informatics and Systems*, vol. 35, 100712, 2022, doi: 10.1016/j.suscom.2022.100712.

[15] W. Shen, X. Wang, and G. Zhou, "Energy and carbon footprint minimization for cloud data centers using hybrid reinforcement learning," *Journal of Cleaner Production*, vol. 323, 130194, 2022, doi: 10.1016/j.jclepro.2021.130194.

[16] R. Singh, V. Garg, and A. Kumar, "A carbon emission-aware virtual machine consolidation strategy for green cloud data centers," *Sustainable Computing: Informatics and Systems*, vol. 39, 100820, 2024, doi: 10.1016/j.suscom.2024.100820.

[17] Y. Wang, H. Zhang, and X. Chen, "Optimizing energy consumption in cloud data centers using deep reinforcement learning," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 2, pp. 320–333, 2021, doi: 10.1109/TSUSC.2021.3076545.

[18] X. Wu, Z. Liu, and W. Hu, "Multi-agent reinforcement learning-based energy-aware scheduling in cloud data centers," *IEEE Transactions on Cloud Computing*, vol. 11, no. 1, pp. 15–29, 2023, doi: 10.1109/TCC.2022.3076546.

[19] R. Yadav, W. Zhang, and Y.-C. Tian, "Energy and carbon optimization for sustainable cloud data centers using reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no.

5, pp. 4113–4125, 2023, doi: 10.1109/TII.2022.3054321.

[20] Y. Zhang, M. Zhao, and S. Li, "Energy optimization for virtual machine placement using DRL in cloud computing," *Cluster Computing*, vol. 25, no. 2, pp. 925–938, 2022, doi: 10.1007/s10586-021-03235-y.

[21] F. Zhao, X. Liu, and S. Li, "A carbon-aware resource allocation framework for cloud data centers using deep learning," *IEEE Transactions on Cloud Computing*, vol. 8, no. 4, pp. 839–850, 2022, doi: 10.1109/TCC.2021.3076547.

[22] P. Zhou, L. Zhang, and M. Li, "Dynamic energy management for cloud data centers: A reinforcement learning approach," *Journal of Network and Computer Applications*, vol. 186, 103097, 2021, doi: 10.1016/j.jnca.2021.103097.