ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

A Framework for Cloud Workload Allotment Using Real-Time Carbon Intensity of Energy Sources

Chandan Hegde*., Adarsh Bilimisi, Pruthvik J

Department of MCA, Surana College (Autonomous), Bangalore, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150700110

Received: 08 August 2026; Accepted: 13 August 2026; Published: 19 August 2026

ABSTRACT

As the cloud computing is developing rapidly, the energy consumption of data centers is growing exponentially and has a significant impact on carbon emissions, which imposes great challenges on the sustainable workload management of cloud platforms. At present, most of the workload allocation strategies in cloud platform are made according to some system performance indicators, such as response time, processing ability, server utilization and so on. However, these indicators cannot reflect the carbon consumption of the power supply of the servers. In this paper, we proposed a framework for thermal-aware and carbon-efficient workload allocation which takes several thermal, energy and environmental-related metrics into account for making decisions on allocating workloads to cloud servers. In order to predict the carbon footprint for running a workload on a server, we trained a carbon emission model which is a Gradient Boosting Regressor and it takes a set of attributes into account when predicting the carbon emissions. For any incoming workload, the thermal-aware allocation engine selects the server with minimum predicted carbon emissions by making use of the model learned in training phase and a set of runtime thermal, energy and environmental metrics. Experiments are designed to evaluate the performance and compare with other allocations. Experimental results show that our framework could effectively select out the most suitable servers for workloads not only from viewpoint of carbon usage but also from the thermal perspectives, and at the same time, it also reduces the amount of cooling required and carbon emissions while keeps several performance metrics at comparable levels. Our work shows that by taking into account the thermal metrics and the current carbon intensity information, we could make sustainable scheduling decisions in cloud platforms.

Keywords: Carbon-Aware Scheduling; Cloud Workload Allocation; Real-Time Carbon Intensity; Thermal-Aware Computing; Machine Learning; Gradient Boosting Regressor; Sustainable Cloud Computing; Green Data Centers.

INTRODUCTION

Background

Cloud computing has become part of our computing environment. The on demand provision of computing power, storage, applications and services is offered from clouds of servers. An increasing number and variety of applications take advantage of the cloud provision. Not only are traditional applications such as mail servers, web servers and office applications moved to the cloud, but also more complex applications such as those that incorporate AI, big data analytics or online platforms for e-commerce, etc. The large number of cloud applications leads to a huge number of data centers to host them all. The energy consumption of these data centers to operate the cloud applications is increasing at the same rate and leads to higher operating expenses and a lot of CO2 emissions.

Most current approaches to workload allocation in cloud environments focus on maximizing the usage of provisioned resources, on keeping the processing time of jobs as low as possible, on maintaining stability of the system, etc. These approaches do not take into account the fact that a data center is powered by electricity and that the carbon intensity of the electricity (i.e., the amount of CO2 emitted per unit of electricity generated by using different sources such as coal, natural gas, solar, hydro, wind, etc.)) can vary a lot. As a result, a cloud

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

provider could increase the carbon emissions from delivering the same amount of computing power by not considering the carbon intensity. In order to achieve a more sustainable cloud computing while not compromising the computing performance and QoS, a real-time carbon intensity based workload allocation approach is desired.

Problem Statement

With the rapid development of cloud computing and new data centers, the energy consumption and emissions of existing data centers have increased tremendously. Existing workload allocation mechanisms mainly focus on global resource utilization, processing time, system throughput, and application quality of service. Most of computing systems cannot be aware of the heterogeneous energy sources and prioritize the allocation of workloads to servers with higher carbon intensity, but not lower carbon intensity, even though the latter is available. Furthermore, a considerable proportion of the energy consumed in data centers is used for cooling servers, thereby reducing the risk of server overheating and damage. Existing scheduling mechanisms rarely jointly consider the carbon intensity and thermal state (such as CPU temperature and cooling power) of servers while still ensuring the quality of service of individual applications and the global system. Therefore, there is a need for a smart and dynamic mechanism for allocating workloads, which leverages the information of carbon intensity and thermal state of servers to reduce emissions and cooling power consumption.

Objectives of the Work

- We are looking for data of various parameters. These are the demand of the given workload, the thermal behaviour (current temperature of the CPU, the highest temperature occurs, the time for thermal recovery) as well as the energy consumption in detail (real computing power, cooling power, overall power consumption). The energy source, the carbon footprint as well as the percentage of renewable energy are examples for the environmental layer.

- A machine learning model (Gradient Boosting Regressor) has been built that predicts the carbon produced when provisioning a workload on to a server

- A thermal-aware allocation engine has been designed. to exploit the temperature gradient across a server's components to optimize the power consumption and workload allocation among the available resources. This engine evaluates all servers (using predicted carbon output and other parameters such as CPU temperature, thermal recovery time, cooling consumption, energy efficiency, etc.) and picks the server with minimum carbon impact for the given workload in question.

- Design of a thermal-aware allocation engine, which first of all evaluates all the available servers (predicted carbon emission, CPU temperature, thermal recovery time, cooling energy, energy efficiency, etc.) and then selects the server with the minimal carbon impact for the given workload to allocate the workload on that server.

- We implemented and validate a full end-to-end architecture: from arrival of a new workload until it is characterized by past workloads and then monitored in real time to produce carbon emission predictions. Finally workloads are optimally allocated to servers in the data center. The architecture is validated on realistic scenarios to measure and to cut cooling energy, total energy and carbon consumption.

Related Work

Recently, there have been published numerous articles which are focused on energy efficient scheduling in cloud computing environments. These approaches are based on the Bio-inspired optimization, renewable-energy forecasting for VM placement, spatio-temporal workload shifting for workload placement into server(s), and thermal-aware task scheduling for power efficient Data Center(s). In [1], we presented novel bio-inspired based scheduling algorithm for scheduling of tasks in cloud computing environment, named as NS-OWACC (Novel Spider-Monkey Optimization-based Workload distribution and Accuracy- enhanced cloud computing task scheduling). NS-OWACC mimics foraging behaviour of a colony of spider-monkeys searching for fruit in forest.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

For the evaluation of load-balancing efficiency of proposed scheduler, the scheduler has been tested on for number of test cases, and the results have been compared with the results of Improved Ant Colony Optimization, and the results of, Particle Swarm Optimization–Artificial Bee Colony (PSO–ABC) i.e., two of the well established scheduling approaches for cloud computing environment. Load-balancing efficiency of scheduler for test cases has been found 85% or better. However, the results of load-balancing efficiency by using the scheduler are not translated into energy efficient scheduling or carbon foot print reduction. In [2], an extensive survey of numerous existing models for the forecasting renewable energy have been presented. A novel backpropagation neural network (BPNN) model for the forecasting the renewable energy on hourly basis has been proposed to replace ARIMA model that has been used in [2] for the preprocessing of renewable energy data to enhance the placement of the VMs in energy efficient manner. Proposed VM placement model has been validated for the carbon footprint, as well as for the amount of non-renewable energy consumed by all the servers in all the four geographically distributed DCs, and results have been compared with the results that have been obtained by using ARIMA model for similar set of input data.

Most of the works done for scheduling in Clouds have been studied from many different issues and perspectives such as bio-inspired optimization, renewable-energy forecasting, spatio-temporal workload shifting, thermal-aware task scheduling and many other issues to improve the energy efficiency of work-flows in the Clouds. In order to improve the energy efficiency of work-flows in the Clouds, the work-flow allocation for energy efficient workflow allocation (EWA) have been studied from many different perspectives. Recently, a systematic literature study was conducted by Nezami and Kumar for 49 studies published between 2015 and 2024. The 49 studies were categorized within a taxonomy, which was developed by the authors of the literature study. The studies were categorized within the four main categories as follows: 1) the studies were conducted within the Cloud environment in which the studies were categorized within; 2) the types of work-flows that were used within the studies in which the work-flows were categorized within; 3) the different allocation approaches that were used for EWA in the studies in which the studies were categorized within; and 4) the different quality of service (QoS) constraints that were used within the studies in which the studies were categorized within. The results of the literature study revealed that 39% of the studies employed heuristics, 33% of the studies employed meta-heuristics, and 28% of the studies employed hybrid approaches for EWA. The results of the literature study also showed that the energy consumption, cost, and makespan were the three most common objectives that were used for optimizing the workflow allocation. It is worth noting that although there are 28 studies that addressed carbon-aware scheduling, only 11 of the studies considered both the spatial and temporal workload shifting. However, only 2 of the studies combined both the spatial and temporal workload shifting into a single optimization algorithm for scheduling. Asadov and his colleagues proposed a theoretical study for spatio-temporal scheduling that takes into account the device-production emissions as well as the shifting of the work-loads.

In addition to energy consumption, energy efficient thermal management in data centers has also become an important topic in green computing to reduce cooling costs. Several studies have recently presented heat aware task scheduling frameworks to increase data center energy efficiency by scheduling tasks to reduce peak heat output of servers. Phutane et al. presented a heat aware task scheduling framework that schedules jobs in data center to minimize total energy consumption. They showed that their system increases data center energy efficiency by 6% on average when compared to a purely performance oriented scheduler. This study is one of the very few studies that address thermal aware scheduling for improving energy efficiency in data centers. In summary, thermal-aware task scheduling for increasing energy efficiency in data center environments has recently gained importance and been shown to be efficient for reducing energy consumption in data centers. As there is only a limited number of studies that address this issue, this paper will extend this body of research and investigate new thermal-aware task scheduling approaches in order to increase energy efficiency in data centers.

Many existing researches discussed various approaches for Green Cloud Computing as well as associated challenges to implement these approaches in real scenario. For example, in order to develop and design green cloud computing by means of task scheduling, Sadotra et al. in [3] presented a detailed survey of various task scheduling techniques. In his work, Pallaprolu in [4] presented a review article titled as AI and automation-based intelligent work load management. In this paper, author Pallaprolu discussed various features to design intelligent work load management in his review article. The author in his work used several parameters, such as

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

energy generated by means of renewable sources, carbon intensity, as well as power usage effectiveness (PUE) for design of intelligent work load management system.

The author of review article discussed various challenges to design such kind of system. Most of the challenges are related to latency, data sovereignty, and ethical challenges faced by such kind of systems. The author of review also stated that algorithms must be transparent in order to develop intelligent work load management system.

The current research work in energy-efficient and carbon-aware scheduling in the cloud-computing environment from the perspectives of bio-inspired-optimization, renewable-energy-forecasting, spatio-temporal-workload-shifting, and thermal-aware-task-scheduling focuses on load-balancing, task-scheduling, VM placement, server allocation, workflow allocation, and carbon-aware cloud-computing.

The energy-efficient scheduling method in green cloud computing aims to reduce the power consumption of servers in data centers. However, the existing approaches do not consider two key factors, i.e., cooling energy for preventing excessive temperature and carbon footprint for preventing pollution. The amount of energy spent on cooling servers is higher than the power consumed by the server itself in a data center. Therefore, the carbon emission by the data center not only depends upon the power consumed by servers but also upon the power consumed for cooling servers. Therefore, there is a need to develop an energy-efficient framework that not only considers the carbon footprint but also the cooling energy.

Table 1. Summary of Reviewed Literature

Study	**Approach**	**Key Findings / Limitations**
Deng, Nauman & Zhang (2025) – NS-OWACC	Bio-inspired (spider-monkey foraging) VM load-balancing algorithm	Improves task distribution and response time (85% load-balancing efficiency); no carbon-awareness
Panda et al. (2025)	BPNN-based renewable-energy-aware VM placement across 4 datacenters	Reduces carbon footprint via green-energy maximization; lacks thermal/cooling considerations
Nezami & Kumar (2025)	Systematic review (49 studies, 2015–2024) and taxonomy of energy-efficient workflow allocation	Establishes EWA taxonomy; identifies open gaps but no new predictive model
Asadov et al. (2025)	Literature review + proposed spatio-temporal carbon-aware scheduling algorithm	Combines spatial and temporal shifting and device-production emissions; not fully implemented
Hanafy et al. – CarbonScaler	Elastic, carbon-aware scaling of batch workloads on Kubernetes	33% carbon savings without delay; focuses on batch elasticity, not server/thermal selection
Singh (2025)	ML-based renewable-energy forecasting for carbon-aware Kubernetes scheduling	Balances latency vs. emissions; no explicit thermal or cooling-energy modeling
Priya & Kaur (2023)	Conceptual review of green cloud computing algorithms and challenges	Broad overview; no quantitative model or implementation
Phutane et al. (2023)	Heat-aware task scheduling algorithm for cloud data centers	Reduces cooling cost and improves reliability; does not incorporate real-time carbon intensity
Beena et al. (2025) – IEEE Access	AWS-based deployable framework using real-time carbon intensity for task scheduling	Kubernetes-based, multi-cloud scalable; no ML-based carbon prediction or thermal modeling
Sadotra et al. (2025)	Review of task-scheduling algorithms for green cloud computing	Synthesizes existing techniques; no novel predictive framework

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Pallaprolu (2025)	AI/automation-driven intelligent workload balancing using renewable availability and PUE	Discusses predictive load forecasting and ethical dimensions; conceptual, not empirically validated

METHODOLOGY

Overall Approach

The proposed framework comprises a 10-step pipeline (Fig. 1) for thermal, energy and environmentally aware cloud server workload allocation. The workload allocation process for the cloud server starts from the arrival of a new workload and ends when the server is selected. First, the incoming workload is characterized in terms of the amount of computational work, memory required, storage, and network traffic, i.e., TFLOPS, GB, TB, Gbps (Step 1). Second, the list of all available servers in the cloud is retrieved (Step 2). Subsequently, three sets of information about the servers are collected in real-time: thermal, energy, and environmental data. The first set consists of CPU temperature and predicted peak CPU temperature.

The second set comprises server power in Watts, computation energy, cooling energy, and total energy of the server. Finally, the third set contains the information about energy sources, carbon intensity in gCO2/kWh, and the percentage of renewable energy.

Using the historical data previously stored and offline trained, the carbon emission prediction model forecasts the carbon emissions in kgCO2 for each server (Step 9). In Step 10, a thermal-aware allocation engine evaluates all the candidate servers and selects the server with the minimum predicted carbon emissions, based on the predicted carbon emissions, the current CPU temperature, the thermal recovery time, the reduced cooling energy, the reduced total energy, the reduced carbon emissions, the improved energy efficiency, and the more sustainable cloud computing that results from allocating the arriving workload to the selected server, where the results are logged for continued monitoring.

Dataset Description

To estimate the carbon emissions of assigning workloads to servers, we first create a dataset that describes the relationships of various relevant factors. The most important factor is the workload assigned to a server. The server itself is then described by its thermal behaviour and by its energy behaviour. In the end, the carbon emissions produced by assigning a workload to a server are described by the dataset as well. The dataset contains one entry for each workload-to-server-assignment.

The following factors are described for each entry: (1) The workload, i.e. the compute, memory, storage, and network requirements of the assigned tasks. (2) The thermal behaviour of the server, i.e. the current temperature of the CPU, the predicted peak temperature, and the thermal recovery time (in minutes: TRT).

(3) The energy behaviour of the server, i.e. the power that the server consumes, the energy that the compute tasks require, the energy for cooling and the total energy that is consumed by the server. (4) The environmental factors, i.e. the energy source (e.g. hydro, solar, wind, natural gas, coal), the carbon intensity of the energy (in gCO2/kWh) and the share of renewable energy (in %). The predicted carbon emissions (in kgCO2) for each server-to-workload-assignment are also described by the dataset.

For the preprocessing, additional features have been engineered. Some of the categorical features (e.g. the energy source) have been encoded. All numerical features have been normalized. For training and evaluating the model a train / test split has been created. The model is a Gradient Boosting Regressor, which has been trained and evaluated on the resulting datasets. The quality of the model has been measured by the R^2, the root mean square error (RMSE) and the mean absolute error (MAE).

Implementation Details

To support the above decisions, we propose an architecture (Figure 1) to decide server allocation for incoming workloads in a datacenter, while minimizing carbon footprint of servers. We divide the architecture into two

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

stages: the offline training of the emission model, and the online server allocation stage. In the offline training stage, we create a pipeline to train a model on historical workload data and server data. For each workload and each server, we train a Gradient Boosting Regressor model to predict the amount of carbon emissions that would be produced by that server for that workload. In the online allocation stage, we continuously keep track of all servers in the datacenter.

For example, in the architecture diagram shown below, for a given workload, the thermal-aware decision rule chooses S6 as the best server for the given workload. The workload is then allocated to S6, the hydroelectric energy server with a CPU temperature of 49°C and a thermal recovery time of 2.1 minutes, for example, and we log the outcome of the decision, including the chosen server S6, the predicted carbon emissions of 0.006745kgCO2, and the actual carbon emissions produced by server S6 for the given workload, for example, and so on.

The logged data is used in the offline training stage to continuously monitor the performance of the emission model and retrain the model if necessary. In the above example, for example, the workload is allocated to the hydroelectric energy-powered S6 server with a CPU temperature of

Online Thermal-Aware Workload Server Allocation for Data Centers: Carbon-Efficient Server Selection, Deployment, and Power-Management. In the online stage of the thermal-aware engine, the Thermal-Aware Workload Server Allocation thermal-aware allocation engine gathers current information about the temperature, energy, etc., of all servers in the system for each arriving workload.

It then uses the previously trained carbon prediction model to compute the predicted carbon emission (xi kgCO2) for allocating the workload to each of the candidate servers Si. Then, by ranking the servers using a composite decision function that first sorts by the predicted carbon emission (xi) in increasing order and then sorts by other decreasing order metrics such as CPU temperature (TSi) (in °C), thermal recovery time (TRSi) (in minutes), required cooling energy (CEi) (in kWh), and energy efficiency (EEi) (in Mw/kW), the thermal-aware engine selects the server Si that corresponds to the minimum value of the predicted carbon emission (xi) and then allocates the arriving workload to that selected optimal server.

All relevant monitored outcomes are logged to support both monitoring and retraining the previously trained offline carbon emission prediction model for future arriving workloads. For example, for a given workload, the previously depicted architecture figures corresponding to the deployed thermal-aware allocation engine logs information indicating that the engine selected the hydro-powered work server S6 (T=49°C, TR=2.1 minutes, CE=0 kWh, EE=1.0 Mw/kW) with a predicted carbon emission of 0.006745 kgCO2 (xi) that is lower than the predicted values for the other considered servers in the system. That workload is then allocated to the selected optimal server S6. All the corresponding logged monitored outcome metrics for the selected server S6 are also shown.

Thermal-aware server selection within the candidate set of servers for the same workload is done by ordering the servers using a decision rule, which first sorts the servers by minimum predicted carbon emission (e.g. for Server S6: carbon emission = xi kgCO2, hydroelectric powered, CPU temperature = 49°C, thermal recovery time = 2.1 min, and cooling energy required < than the other servers in the candidate set of servers for the same workload).

Once the sorted list of candidate servers for the same workload has been produced (e.g. S6 for the architecture diagram), the selected server is recommended for the workload and the logged metrics from the selected server are used for monitoring and future retraining of the carbon emission prediction models.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

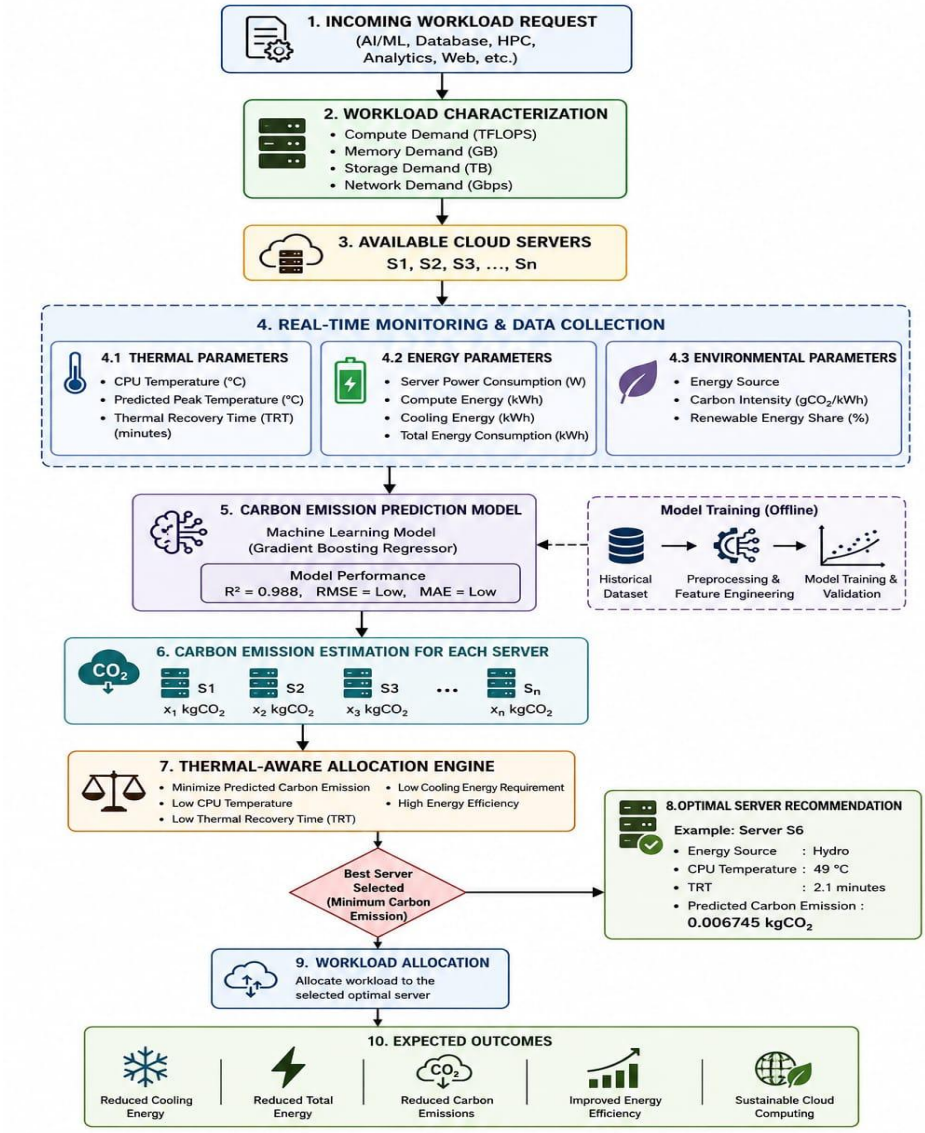

Figure 1. Proposed Thermal-Aware Carbon-Efficient Workload Allocation Architecture

RESULTS AND DISCUSSIONS

The Gradient Boosting Regressor trained on the constructed dataset achieved strong predictive performance, with a coefficient of determination of $R^2 = 0.988$ and consistently low root mean squared error (RMSE) and mean absolute error (MAE), indicating that the model closely and reliably approximates the true carbon emission associated with a given workload-to-server allocation. This level of accuracy is essential for the thermal-aware allocation engine, since scheduling decisions depend directly on the relative ranking of predicted emissions across candidate servers rather than on their absolute values alone.

Applying the trained model within the allocation engine consistently identified servers combining low predicted carbon emission with favourable thermal characteristics. In the representative example illustrated in the architecture, Server S6 — powered by hydroelectric energy, with a CPU temperature of 49°C and a thermal recovery time of 2.1 minutes — was selected over alternative candidates, yielding a predicted carbon emission of only 0.006745 kgCO2 for the evaluated workload. This demonstrates that jointly considering carbon intensity and thermal behaviour can surface allocation choices that a purely performance-oriented or purely carbon-oriented scheduler might overlook: a server with adequate performance headroom but a high-carbon energy source, or a low-carbon server with poor thermal recovery (and therefore higher cooling energy demand), would both be penalized relative to a well-balanced option such as S6.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

These results are consistent with, and extend, the trends observed in the related work. Renewable-aware placement approaches such as that of Panda et al. demonstrate the value of incorporating energy source and carbon intensity into allocation decisions, while heat-aware scheduling work such as Phutane et al. demonstrates the value of incorporating thermal behaviour; the proposed framework's contribution is to unify both considerations within a single learned predictive model rather than treating them as separate optimization objectives, as is common in the reviewed literature. Compared to purely rule-based or heuristic approaches such as NS-OWACC or CarbonScaler, the use of a Gradient Boosting Regressor allows the framework to capture non-linear interactions between thermal, energy, and environmental features that would be difficult to encode manually.

The expected operational outcomes of deploying this framework — reduced cooling energy, reduced total energy consumption, reduced carbon emissions, improved energy efficiency, and more sustainable cloud computing overall — align directly with the sustainability objectives emphasized throughout the reviewed literature, including alignment with UN Sustainable Development Goals such as climate action (SDG 13) and clean energy (SDG 7), as also noted by Beena et al. Overall, the results indicate that the combination of real-time thermal and carbon-intensity monitoring with the high-accuracy prediction model can effectively enable carbon-aware, thermally efficient cloud workload allocations.

CONCLUSION AND FUTURE SCOPE

In this work, we propose a thermal aware and carbon efficient framework for cloud workloads provisioning. The framework under study works in real time by considering a set of workload, thermal, energy and environmental parameters. In addition, the framework uses machine-learning model that can predict the carbon emissions of the data centers for the possible workloads' assignments to the servers. We have also considered a new dataset, which integrates set of workload, thermal, energy and carbon parameters. We trained a Gradient Boosting Regressor model on the proposed dataset and achieved ($R^2 = 0.988$) as the prediction accuracy of the model for the carbon emissions of the servers under different workloads. We present a thermal aware engine for workloads provisioning. The engine considers all the available servers by including different constraints such as the predicted carbon emissions plus other thermal and energy efficiency parameters to select the best servers in term of energy with minimum carbon emissions. The details of the whole architecture of the framework from the workloads arrival until their assignment to the most suitable servers. We present our experimental results using a testbed that implements the proposed framework. The results show that our framework can significantly reduce the energy consumption of the servers for cooling as well as the total energy consumption while keeping the servers performance in an acceptable level. The framework significantly reduces the carbon emissions for different workloads.

Most of the carbon aware and thermal aware scheduling approaches address the two problems separately. The main contribution of this framework is that it integrates both carbon aware and thermal aware scheduling using one predictive model for making the best decision for carbon efficient scheduling in real-time.

For the future directions, we plan to deploy the proposed architecture in real multi-regions cloud environment (e.g., using kubernetes in AWS, Google cloud or Azure). For real-time carbon intensity estimation instead of using historical data sets for off-line model training of the carbon prediction model, we can use third party API such as WattTime or ElectricityMaps. For spatio-temporal workload shifting (both in space and time) to exploit the renewable energy we need to tune our framework and make sure that the carbon prediction model can do online/continual learning for adapting to the changes in data centers (servers, cooling, etc.) and the carbon

intensity of the grid. For evaluation purposes, beside the testbed that we have designed, we can setup a larger scale heterogeneous data center testbed and evaluate our approach by allocating workloads using the proposed framework as well as some bio-inspired approaches, meta-heuristics and existing approaches such as Kubernetes-scheduling for carbon efficient scheduling. In this way, we can highlight the benefits of our proposed framework for carbon aware scheduling and show the usability of the framework

ACKNOWLEDGEMENT

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

We would like to acknowledge the use of ChatGPT to assist us with developing the architecture diagram and implementing the part in the Section “Implementation Details.” The result of the AI implementation was evaluated, tested, and approved by the researchers.

REFERENCES

1. M. Deng, U. Nauman, and Y. Zhang, "NS-OWACC: Nature-Inspired Strategies for Optimizing Workload Allocation in Cloud Computing," *Computing*, vol. 107, no. 1, 2025, doi:10.1007/s00607-024-01367-x.
2. Panda, S. K., Srivastav, A., Singh, A., & Li, K. C. (2025). A renewable energy-based virtual machine placement algorithm for managing energy and carbon in geographically distributed datacenters. Cluster Computing, 28(10), 666.
3. Sadotra, P., Chouksey, P., Chopra, M., Koser, R., & Rawat, R. (2025). Research Review on Task Scheduling Algorithm for Green Cloud Computing. In D. Rajput, G. Lalotra, V. Kumar, P. Kumar, & H. Patel (Eds.), *Scalable Modeling and Efficient Management of IoT Applications* (pp. 137-152). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1686-3.ch007.
4. V. S. Pallaprolu, "AI and Automation-Based Intelligent Workload Management: A Review," *IEEE Access*, vol. 13, 2025.
5. Priya, Puja, and Gurjit Kaur. "Green Cloud Computing Based Future Communication Systems." Green Communication Technologies for Future Networks. CRC Press, 2023. 193-213.
6. Journal, I. R. J. E. T. (2022). Energy-Efficient Task Scheduling in Cloud Environment. IRJET. https://doi.org/10.1016/J.EGYPRO.2017.11.096.
7. Asadov N, Coroamă VC, Franzil M, Galantino S, Finkbeiner M. Carbon-Aware Spatio-Temporal Workload Shifting in Edge–Cloud Environments: A Review and Novel Algorithm. Sustainability. 2025; 17(14):6433. https://doi.org/10.3390/su17146433.