Achieving Green Data Centres for Sustainable Cloud Computing

MSc Research Project
Master of Science in Cloud Computing Information

Mariam Ishaq
Student ID: X23259817

School of Computing
National College of Ireland

Supervisor: Sean Heeney

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mariam Ishaq
Student ID:	X23259817
Programme:	Master of Science in Cloud Computing Information
Year:	2024
Module:	MSc Research Project
Supervisor:	Sean Heeney
Submission Due Date:	12/12/2024
Project Title:	Achieving Green Data Centres for Sustainable Cloud Computing
Word Count:	6435
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Achieving Green Data Centres for Sustainable Cloud Computing

Mariam Ishaq
X23259817

Abstract

The growth of cloud computing comes with a cost, data centres, extremely power hungry and large CO2 producers, are becoming increased sustainability problems. Their electricity consumption might emerge at 3000 TWh annually by 2030 according to estimates, or 8000 TWh in wors-case scenario. This research explores how data centre energy consumption can be minimized to extend cloud reliability and efficiency.

Using Google Cluster Workload Traces, derived from real-world dynamic workloads in Google's Borg clusters, the research examines computational resource utilization and energy profiling. Machine Learning models - GBDT, DQN, and HEPGA are used to estimate energy consumption in a cloud computing data centre to optimize on resource utilization and optimal resource scheduling. Simulations of energy consumption, conducted using Python's SimPy framework, demonstrate energy-saving potential under various workload conditions.

The results highlight the power of predictive modelling with low MSE in forecasting energy demands and the impact of dynamic scheduling algorithms on reducing consumption. This approach demonstrates the path towards green and high-performance data centres for the next generation meeting the ecological issues of cloud computing.

1 Introduction

Data in the past couple of decades has grown exponentially, along with that the number of Data Centers has risen. Data Centres are the most critical component, or we can say, the backbone for Cloud Computing. In our generation, sustainability has been the number one priority for all the industries along with Cloud Computing. Unfortunately, Data centres consume massive amounts of energy causing significant co2 emission. According to IEA, 330 megatons of co2 emission was generated by data centers in 2020, globally. Researchers all over the world, trying to find revolutionary techniques to cut down the carbon emissions from Data Centres.
The total metered electricity consumption by data centers as a percentage of total consumption for the entire time span from 2015 to 2022 is depicted in this bar chart below. An analysis of the data reveals that the electricity consumption by data centres is on an appreciating trend during this period, proving that they are asserting a more incrementing utility of energy and hence inflicting more harm to the environment.

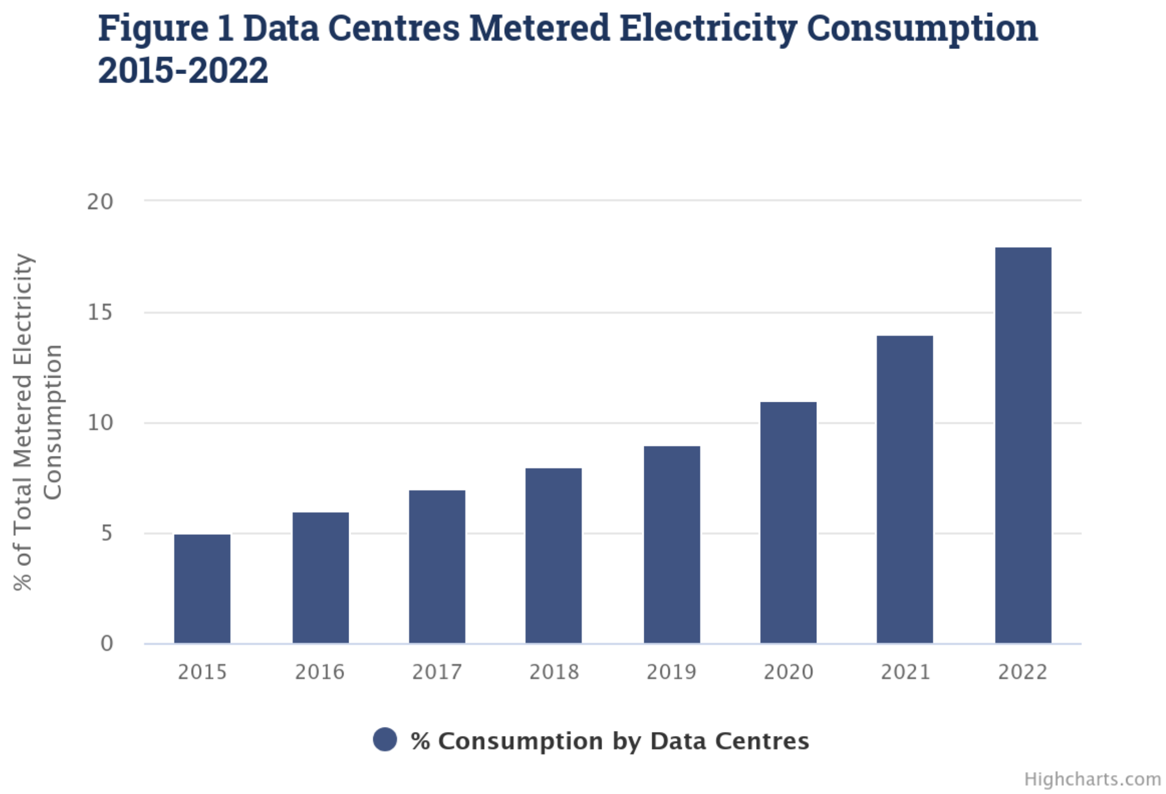

Figure 1: Data Centre Electricity Consumption in Ireland by Year (Source: cso.ie)

2015–2017: It reveals a relatively slow and systematic increase, which also points to the slow but sure growth in data center activities and powers.

2018–2020: There is a sharper increase, presumably as a result of systematic-digitisation, growing embrace of cloud-computing and ever-heavier reliance on data driven tools.

2021–2022: The largest increase is in the year 2022 in which data centers used close to about twenty percent of electricity. This might have been occasioned by rising trends of artificial intelligence, big data analytics and increasing needs of data storage.

Such a tendency proves the need for designing more efficient data center solutions from the environmental impact perspective, although there is increased demand for computational power.

Main challenges towards achieving green data centres are, the fast and continuous need of data availability. Cloud-Computing is used by several critical industries such as healthcare, financial services, energy and utility, and any operational inefficiency in their cloud can cause severe damage to security, property, and life. This dependency on data centres makes it hard for researchers to find a perfect solution that is sustainable and efficient at the same time.

The importance of this research is maintaining high levels of technical efficiency along with cutting down the energy need resulting in co2 emission. We want to contribute to the creation of next generation green data centres through this research, as the cloud computing industry expands in size and capacity along with its need for more data centres.

The research question that motivates us to dig deeper into this, would be – **How can we achieve energy efficiency in data centers for cloud computing, without affecting the cloud performance?** This question motivated me to find out the best techniques to get energy efficiency in Data Centres, exploring advanced cooling techniques, energy efficient hardware, software-based scheduling, and renewable energy sources.

This research leverages a dataset with 400,000+ entries and 34 columns, from google cluster workload traces dataset 2019, from Kaggle, for this research. This data provides 8 different clusters from Google's Borg compute clusters for the entire month of May (2019). These clusters are responsible for job submissions, job scheduling and resource usage across huge cloud computing workloads. Main reason behind choosing this dataset is, it consists of real-world dynamic workloads, so we can accurately analyze the resource consumption to schedule it efficiently and reduce energy usage. It is understandable that there are some limitations to the dataset, such as not having user data for operational priorities, along with its huge size & complexity. Also, we only have data for May, which restricts us from finding out how the pattern changes during different seasons. The dataset does not directly provide energy usage statistics, so feature-engineering 'estimated power consumption' column to be calculated for power consumption through resource usage such as 'CPU utilization.'

The structure of this report is as follows -

1. **Abstract** – An overview of the topic is provided.
2. **Introduction** – Background, importance, objective, outcome, and limitations are discussed.
3. **Related Work** – Different researchers on the topics related to our project are discussed and reviewed. This section provided a fantastic opportunity to develop research methodology based on past research.
4. **Methodology** – Discussed research methods
5. **Implementation** – Implementation process is discussed, evaluated the results and discussed how the outcome met the proposal.
6. **Conclusion** – Discusses if objectives are met, how the research question is answered and restates the key findings along with future scope.

2 Related Work

There have been multiple research projects on energy optimization techniques in Data Centers, with the rapid growth of the Cloud Computing industry. Before I get started with my own research, it is important to learn from past research on the topic, to take inspiration and learn about the current industry standard.

Advanced Cooling and AI-Based Scheduling Techniques for Next-Gen Data Centers

Zhang and Shan (2023) in their research showed that next-generation data centres need to include advanced cooling techniques, integrate with renewable energy sources, improve hardware for less energy consumption, and discussed AI-based scheduling techniques. For cooling techniques, authors emphasized outdoor air for cooling instead of mechanical cooling, although they have mentioned more exploration on its availability. They have equally emphasized including renewable energy, and exemplified how Google and Microsoft have incorporated solar, wind and hydroelectric sources for their data centers for their large-scale projects to reduce carbon emissions. In the next part, authors focus on AI-based scheduling techniques, important for our project. According to the study, scheduling resource allocation dynamically as per demand can significantly reduce energy consumption in data centres.

Combining software-based power management with green computing practices

Another great approach by Mark and Bommu (2024), discussed the problems of lowering the carbon footprint in cloud data transmission. Their suggestion includes routing protocols optimization and the use of caching to reduce energy consumption. Just as our project targets the hindrance to green computing, this paper's aim is also to reduce co2 emissions across all components of cloud infrastructure. This paper has emphasized on integrating renewable energy sources along with advanced cooling systems and its effect on reducing the overall power consumption in data centres. They have throughout mentioned in their studies, the importance of integrating renewable energy sources with better cooling systems as a suggestion for hardware's perspective. For the software side, software-based optimization has been considered by them. Allocating resources dynamically is one of the biggest challenges, and use of machine learning is considered the best method by Mark and Bommu. Their project focuses on bridging the gap by "combining software-based power management with green computing practices to create a holistic energy-efficient data centre solution."

In the article Katal and Dahiya (2023) completely describe the evolution of cloud computing and its effects to the data center including the power usage effectiveness (PUE). The paper presents a general background of the evolution of the internet, the development of hyperscale data centers, an explanation of PUE, and an overview of green IT efforts to address environmental problems posed by data centers. They also present key research questions that form a basis for other subsequent research in green computing, especially software power management, virtualization and impacts of data centers on the environment.

S R Raja (2024) offers an extended understanding of the role of software-based solutions focusing on how SVMs allow one to build more accurate forecasts of workload patterns to improve resource usage. Moreover, CNNs improve continuous control over energy-consuming parts, such as cooling systems, which constitute an essential domain of energy effectiveness. RNNs are well suited to the forecasts of power for the irregular energy needs as compared to LSTMs, which provide a more progressive approach compared to the fixed systems used in energy management.

Thus, by combining these models, the study puts forward a more efficient approach to power management in green data centers, which is more flexible and sensitive as compared to hardware-based solutions. This work supports the idea of the hybrid of software optimization and the employing of more advanced machine learning, as discussed by Mark and Bommu but also analyze the current issues with both the hardware and software approach to the problem, stating that it is still insufficient for providing sufficient capabilities for the increasing energy needs of cloud infrastructures.

Machine Learning Algorithms for Dynamic Resource Allocation

Balachandar in his research, presents a comprehensive methodology for dynamically allocating resources that is quite relevant and helpful for deciding the best algorithms for our model development. He proposes three machine learning algorithms – Gradient Boosting Decision Trees (GBDT), Deep Q-Network (DQN) and Genetic Algorithm (GA). As a Decision Tree algorithm, GBDT's speciality in classification accuracy and handling diverse data types, makes it an important tool for optimizing resource allocation. GA determines resource strategies for achieving performance objectives, and the Q-learning algorithm of DQN as a reinforcement learning model improves over time in the decision-making for the system's resources dynamically. Through their research, we found that GBDT has significantly better throughput and response time than DQN and GA, making it the most effective algorithm in edge computing contexts. Their research states the importance of machine learning in dynamic resource allocation in edge computing environments, showing us the best algorithms to consider for the purpose.

Reinforcement Learning as an Alternative to Heuristic Scheduling

For reducing operational cost, most of the traditional scheduling techniques in cloud computing environments, uses heuristic-based algorithms, such as first-fit and priority-based approaches. According to Reza and Zhao (2021), this technique has some shortcomings in adapting to diverse resource demands and proved to be quite inefficient in cloud computing due to its varying workloads and need for dynamic memory allocation Reza and Zhao (2021), in their paper, raised concerns about the heuristic-methods and suggested reinforcement learning (RL) as an alternative. They have shown how reinforcement learning (RL) is more efficient in a dynamic workload environment by optimising models for performance-related metrics like job completion time, CPU, or other resource utilisation, considering energy efficiency at the same time.
In their (Reza and Zhao, 2021) model, they showed us the effect of reward signal in performance of RL, and how critical role the design has for training the RL models, because the poorly designed reward signal cannot be as useful in energy efficiency. Their focus on this reward signal is significant for our research, as its role in balancing energy efficiency and optimum performance is our research goal.

SimPy for Simulation and Optimization in Biomass Supply Chains

Pinho and Coelho (2021) in their paper utilized SimPy for distinct-event simulation. They have developed a framework for biomass supply chains optimization, which involves several stakeholders and uncertain situations, quite like our data centre environment, from the point of uncertainties. Their framework coordinates various planning levels like a scheduler but more resilient and flexible because this research integrates machine learning models and simulation to forecasts within a closed loop. Pinho and Coelho (2021) using SimPy for supply chain optimization provides researchers with an example of how to address operational challenges using simulation.

Testing Scenarios and Simulation Techniques for Energy Efficiency

From Mansouri and Ghafari (2020) paper we understand the importance of finding the correct scenarios to test our scheduler. As one of the main motives of this project is to keep the performance unaffected, it is very important to check how the models are working on different workloads. To understand the fluctuations in energy consumption, authors mention compute intensive and data intensive workloads, to see how their model performs, in these various situations. These use-cases are also great ways to strategize the optimization techniques. Different operation situations, such as peak times and off-peak times, as well as use of renewable energy sources are discussed. Also, I have learnt how environmental factors can affect energy efficiency, temperature and humidity contributes to data centre cooling systems, which is an excellent use case to simulate and observe. Failure recovery scenarios for cloud computing components in data centres, related to energy consumption, this is a very crucial scenario to simulate, according to the authors.

They have also considered a range of simulation techniques that we can adapt, such as computational fluid dynamics (CFD) – which helps with simulating air circulation through servers and cooling performance in data centres. Discrete Event Simulation (DES), the most important one for our project, this approach helps simulate operational scenarios for schedulers while energy optimising. Agent-Based Modelling (ABM), on the other hand, helps understand complex interactions between different components and their effect on energy consumption.

Hybrid Enhanced Particle Genetic Algorithm (HEPGA) for Scheduling Optimization

Mikram and Kafhali (2024) discussed Hybrid Enhanced Particle Genetic Algorithm (HEPGA), to optimise scheduler for cloud computing components. Authors focused on performance metrics like Makespan and resource utilization in their study. This study is helpful to guide me through the evaluation of my model. One of the unique aspects of this study is weight sensitivity, authors found out how weight distributions for fitness function significantly changes the model's efficiency. We can understand the importance of parameter tuning from this study. This research shows how simply increasing the resource availability does not decrease Makespan. In fact, task scheduling performed much better when simulating constrained resource availability. This study is a huge benchmark for the cloud computing industry as well as beneficial for our project.

EnergyAware Scheduling and Green Computing Objectives

Of the reviewed literature, the most comprehensive work on energy-aware scheduling in HPC can be found in the article by Author1 and Author2 (20xx) since this field has shifted from purely performance-oriented designs to power-efficient systems. Metrics, heterogeneous systems, and algorithms are covered by the authors, and scheduling solutions including DVFS and power capping are also described. Recognized gaps in algorithms like reinforcement learning, fuzzy logic, and evolutionary techniques of energy efficiency their overview shows deficiency such as the requirement of more effective energy monitor and impulse to energy point tradeoff. Furthermore, the authors call for future research in incorporating machine learning and auto-configurable power capping systems. Nevertheless, realizing that the study has many merits, one might have expected more detailed consideration of the integration of renewable energy sources and workload fluctuations in real-environment conditions.

This, in turn, has been supplemented by the work of Sharma and Joshi (2023) which posits the retrofit subprocesses, outsourcing, and dynamic resource deployment as key factors in the realization of green computing goals' objectives. It correlates well with energy-aware scheduling in the cloud computing data centers and provides information about some greedy hybrid strategies such as weighted round-robin. But their work does not include the evaluation using realistic workloads with a wide variation and less emphasis on faults. Consequently, the research highlights the possibilities of integrating renewable energy into scheduling algorithms which is another open research area. Still, their work gives historical knowledge on employing simulation tools, including MATLAB, and hybrid approaches for enhancing energy efficiency in cloud, which serve as the subsequent reference for researching sustainable solutions and dynamic scheduling applications.

Almutairi and Aslam (2023) take a novel approach by addressing the need for robust communication resources in cloud applications, which are critical for user experience in terms of bandwidth and traffic management. Their proposed Energy and Communication (EC) aware scheduler algorithm minimizes energy consumption and traffic load in green cloud computing. The EC-scheduler leverages a Multi-Objective Leader Salp Swarm (MLSS) for efficient traffic distribution and an Emotional Artificial Neural Network (EANN) for resource allocation. By employing GreenCloud simulation, the authors demonstrate that their approach improves energy efficiency by 37.59%, throughput by 50%, and reduces energy consumption by 34.2%. These enhancements are substantial, but areas for improvement such as how this algorithm would work at very dissimilar levels of workload and how to fully dilute congestion are still unsolved.

Altogether, the above-mentioned works offer a systematic groundwork for enhancing energy-conscientious scheduling and energy-efficient computing. Authors Kocot and Czarnul (2023) have focused well to address the HPC related issues and opportunities, whereas Sharma and Joshi (2023) as well as Almutairi and Aslam (2023) present more insights of cloud computing. Collectively, their work emphasized integration of renewable energy, fault tolerance and sophisticated algorithms for mental and sustainable computing solutions. These works provide the foundation for future investigations – specifically for scaling green computing objectives, accommodating fault tolerance, and promoting hybrid studies on different environments.

3 Methodology

This research proposes a methodology combining energy consumption prediction with an intelligent scheduling system that minimizes energy use without affecting the quality of service. As the system utilizes machine learning and simulation methods to determine how resources should be allocated over a given period, total cloud data center efficiency is made better, energy is conserved and sustainability achieved.

i) Data Collection and Preparation The dataset applied in this study is Google Cluster Workload Traces (2019) that contains information on resource consumption metrics, such as CPU and memory, and workload distribution across Google data clusters. Since direct energy consumption data is not available in the dataset, we leverage the relationship between resource utilization and energy consumption, which has been established in prior studies. For instance, workloads that take high CPU and memory to run are also likely to consume more energy. Energy consumption calculation is based on the power modeling approach commonly used in research. CPU and memory usage values are used as proxies for energy demand.

We calculate approximate energy consumption using these features with a formula derived from existing power consumption studies in data centers. This formula looks like:

$$E = \alpha \times (\text{CPU Utilization}) + \beta \times (\text{Memory Utilization})$$

where α/alpha and β/beta are scaling factors based on the contribution of each resource to overall energy consumption. These factors can be adjusted according to known power consumption models for specific hardware.

ii) Energy Consumption Prediction - Energy consumption is predicted using XGBoost, DQN and GA models. This research used CPU usage and memory usage as performance indicators, as these are more closely correlated with energy consumption in data centers. Equipping the above-mentioned models with these features will predict how much energy will be used for each task.

The steps include pre-processing the data from basic cleaning to data engineering where the additional features such as density and latency of the resources, in addition to CPU and memory usage are incorporated. They allow the model to consider all sorts of other characteristics that could affect energy consumption, other than the bare CPU and/or memory metrics.

To justify the performance of such a model, we apply averages Mean Square Error [MSE] and R- R-squared [R]. MSE quantifies, on average, how much the predicted values deviate from the actual energy consumption, with the use of the squared differences, which makes sense since it gives a clear picture of how well the model has performed with the ability of determining a minimum variance as the best-case scenario. The value of R-squared also enables us to weigh the accuracy of the model since it shows the percentage of the variance to energy consumption. Ideally, lower MSE and a higher R-squared stand for the higher predictive accuracy.

iii) Task Scheduling for Energy Efficiency - The task scheduling algorithm is used by basing the performance on the predicted energy consumption to enable workloads to be scheduled at times when energy consumption is at the lowest while improving the system's performance.

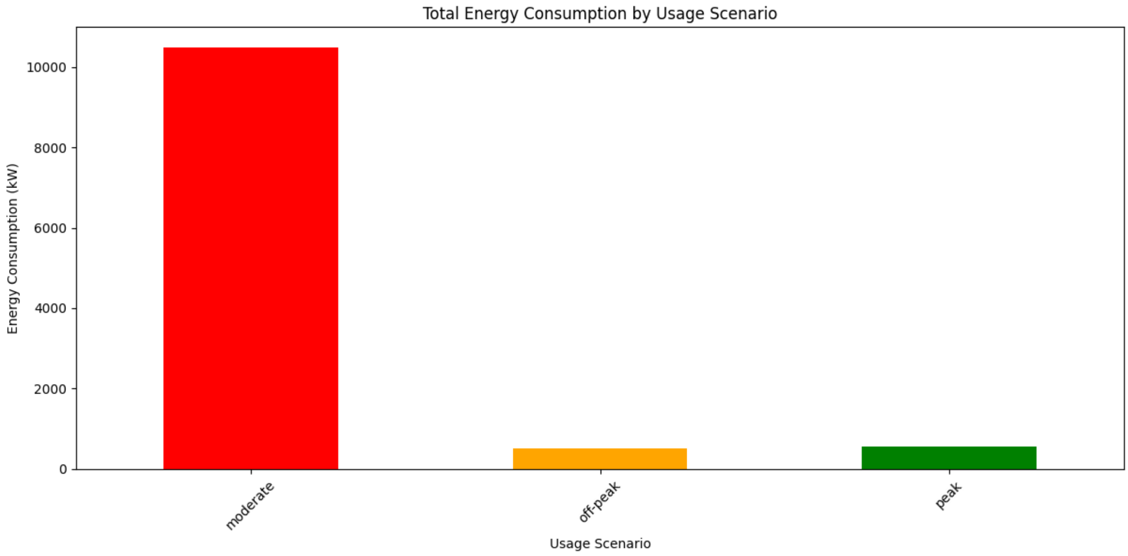

Figure 2: Usage Scenario Distribution based on Energy Consumption

The bar chart visualizes the energy consumption for three distinct task scheduling categories in a cloud data center:
(i)Peak, **(ii)Moderate**, and **(iii)Off-Peak**
.
These categories are derived from the forecast of the resource demands for tasks, and categorized by the scheduling algorithms for improving energy consumption without compromising the performance of the system.

Peak Tasks: These issues require the greatest CPU and memory, meaning that the energy consumption rates are considerably higher. We recall, after analysing the charts above that Peak tasks utilise the most energy among other task scenarios.

Moderate Tasks: These tasks lie mid-range workload categories and do not fit the Peak or Off-Peak categories hence their moderate energy consumption. Their energy consumption is between that of Peak tasks and those of Off-Peak tasks.

Off-Peak Tasks: Activities that can take place at any time and require hardly any resources are planned for low demand time like during evening or on a weekend. This sees a great assurances of lower energy use as evidenced by the lower bar in the chart.

The scheduling algorithm is designed to maximize energy consumption by initiating tasks only during Off-Peak hours, if at all possible, without straining system efficiency or task due dates. The reduction factor used in Off-Peak tasks expect resources to be much available in low demand hours resulting to light utilization of cloud infrastructure.

The system ensures that Off-Peak energy consumption is kept to a minimum, as is wasteful Peak period energy usage. This strategy of power distribution is dynamic in fashion, which means that the power consumed over the various components is dynamic so as to balance power consumption over the various components and yet the performance is uniform.

iv) Simulation and Evaluation the SimPy package is used for cloud simulation, which models a realistic data center environment. To simulate the tasks, resources, and energy consumption in the cloud, Simpy provided the platform to test the effectiveness of the task scheduler in real-time.

During the simulation, tasks are executed based on the scheduling strategy, and energy consumption is calculated both before and after the scheduler's optimization. The goal is to compare the predicted energy consumption (based on workload characteristics and resource usage) with the actual consumption during the simulation, which incorporates the scheduling changes.

This step assists in the evaluation of how the scheduler achieved some level of energy efficiency in the cloud whilst completing more operational tasks, such as task response time and service latency.

Design Specification of the Workflow

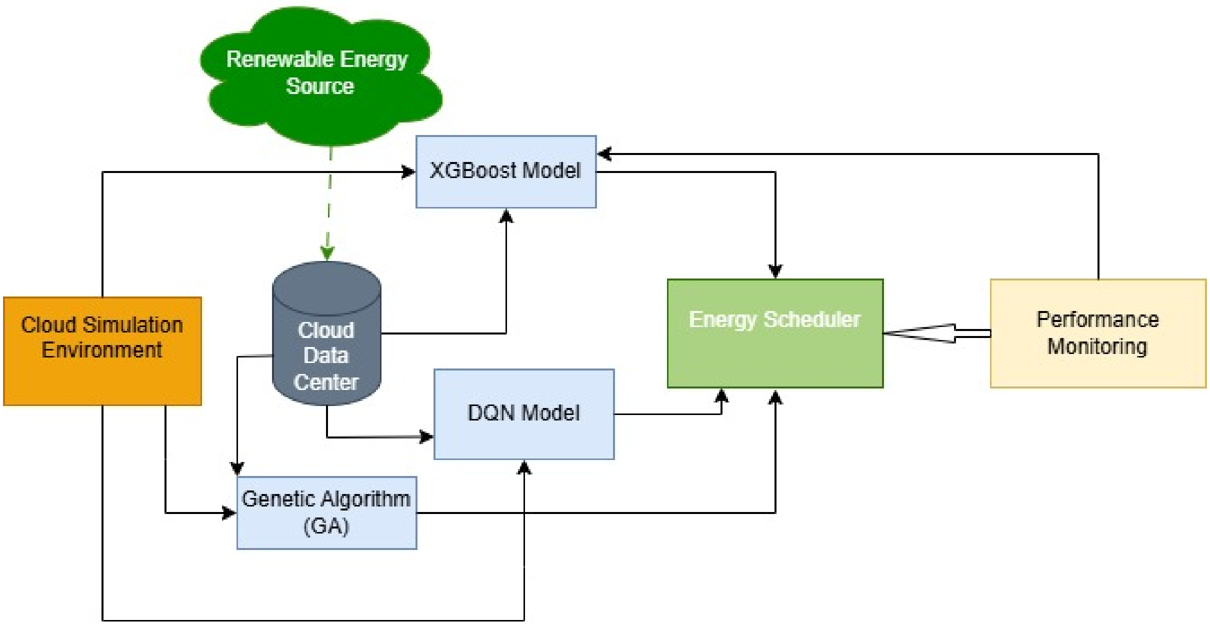

Figure 3: Architecture Diagram of the Workflow

The following diagram provides an overview of the elements that are part of the suggested energy saving system for cloud data centers and efficient energy management, as well as a connection to renewable power sources.
Cloud Simulation Environment (Orange Box): This is the setting for the whole simulation The environment of the simulation The specific setting where the whole simulation occurs It mimics the cloud data center environment and it is where the various optimization initiatives are first deployed.

Cloud Data Center (Gray Cylinder): This is the central site in which the cloud-based tasks are being performed. The data center is an energy-intensive resource and here the idea is to manage the energy usage effectively.
Renewable Energy Source (Green Cloud): This part brings the concept of sustainability into the system by replacing the doubtful energy, be it solar or wind, in the system and thus help in the energy efficient storage of the data center.
XGBoost Model (Blue Box): This model of machine learning is employed to forecast energy consumption in the data center. It is essential when it comes to figuring out when and how energy is used so that organs of major authority can come up with a timetable on when to schedule energy usage.
DQN Model (Blue Box): The basic approach of machine learning is implemented through reinforcement learning using the Deep Q-Network (DQN) model. It learns about the actions that are best suited for minimizing energy consumption while at the same time ensuring adequate performance and quality of the services delivered in the data center.
Genetic Algorithm (GA, Blue Box): Crossover, mutation, and selection models are used in the genetic algorithm to arrive at the best way in which to schedule the tasks. It provides an opportunity to determine the optimal approaches that would be cost-efficient in terms of energy wasting in the context of distribution of workloads.
Energy Scheduler (Green Box): This component utilizes the result of the XGBoost, DQN and GA models to allocate tasks in an efficient and most preferably during the peak energy usage. As a result it ensures that tasks are scheduled for off-peak hours where energy demand and therefore energy prices are still low.
P**erformance Monitoring (Yellow Box):** This block is used to supervise the energy scheduler and the general system performance. It assesses the capability of the scheduler in achieving the energy requirement and the functionality of the system.
Explanation of the Flow
The simulation of the Cloud Data Center operation is provided through the **Cloud Simulation Environment**, which combines several models and strategies.
The **Cloud Data Center** operates over tasks and power. The optimization process emphasizes to whom the task can be delegated so that energy use is lessened.
The **XGBoost Model** deals principally with the study of the consumption of energy and the **DQN Model** focuses on learning of the best schedule of tasks using Q-Learning.
The **Genetic Algorithm**, therefore, continues to enhance the task scheduling to optimize energy consumption additional by adapting the scheduling techniques.
This **energy scheduler** integrates all the models to work on its operations in order to minimize the energy consumption during tasks, especially when renewable energy is on.
Performance Monitoring has the responsibility of analyzing if the system is delivering on the energy saving and performance promises made.
This architecture aims to optimize the scheduling of tasks operations in the cloud data center in order to reduce as much power use as possible while providing quality services in the cloud, perhaps by tapping into renewable energy and smart scheduling models.

Tools and Technologies Used:

Python: Applied to data manipulation and analysis to train the models for energy consumption prediction and to carry out simulations.

XGBoost: A gradient boosting ML model, used for predicting energy consumption based on the CPU and memory loads.

SimPy: A discrete-event simulation library used to model and simulate the cloud data center, evaluate the performance of the task scheduler, and calculate energy savings.

AWS Virtual Machine (EC2): Used for scalable cloud infrastructure to run simulations and handle large datasets efficiently.

4 Implementation

Energy consumption in this study was estimated based on the time taken to complete a task and the resources utilized, including CPU and memory usage. This approach follows established power models where energy consumption is a function of resource usage and task duration. Specifically

CPU Power Consumption: It is estimated using the CPU usage percentage; each usage point (such as 10%, 50%, 100%) is associated with the given power consumption.

Memory Power Consumption: Similarly estimated in relative proportion to memory usage percentage, the higher the usage the more energy to be consumed.

These values were combined to estimate total energy consumption for a given task, using the formula:

$$\textbf{Energy Consumption} = \textbf{CPU Usage} \times \textbf{CPU Power Factor} + \\ \textbf{(Memory Usage} \times \textbf{Memory Power Factor)}$$

This estimated energy consumption for each task was the target variable for the machine learning models.

Three machine learning models were implemented to predict energy consumption and optimize resource usage:

XGBoost: This gradient boosting algorithm was also picked as a suitable algorithm for working with tabular data thereby giving me accurate energy consumption prediction while avoiding over-fitting.

Deep Q-Networks (DQN): A reinforcement learning approach, DQN was used to model dynamic resource allocation policies in cloud systems, leveraging a Multi-Layer Perceptron (MLP) to approximate resource allocation decisions.

Genetic Algorithm (GA)-Optimized Linear Model: This approach used evolutionary methods to optimize parameters for energy consumption prediction, offering a novel approach to resource utilization.

Once energy consumption was predicted, task scheduling was implemented to optimize energy usage while maintaining performance.

The task scheduler was designed to allocate tasks to off-peak hours whenever possible to reduce energy consumption without sacrificing performance. Tasks were classified into two categories based on their resource usage:

Peak Tasks: Tasks with high CPU and memory consumption.

Off-Peak Tasks: Tasks with lower resource usage, where energy consumption could be reduced by up to 10

The scheduler used these classifications to allocate tasks to times when resources were less in demand, thereby minimizing energy consumption while maintaining the reliability of cloud services.

To evaluate the effectiveness of the energy-efficient task scheduler, a simulation was run using the SimPy package. This package modeled a cloud data center environment, accounting for task execution time, resource usage, and energy consumption during both peak and off-peak hours.

Cloud Simulation: SimPy was used to simulate the processing and management of workloads in a cloud data center. The simulation examined the effectiveness of the task scheduling strategy in reducing energy consumption while meeting performance targets.

Results Evaluation: Energy savings were calculated by comparing the predicted energy consumption with the energy consumption achieved through the scheduler. Key metrics like total energy saved and the effectiveness of the scheduling strategy were used to evaluate performance.

Tools and Technologies

Programming Language: Python, with libraries such as Pandas, NumPy, SciPy, and Scikit-learn for data processing and machine learning modeling.
Machine Learning Models: XGBoost for energy consumption prediction, DQN for dynamic resource allocation, and GA for optimization.
Simulation: SimPy for modeling cloud data center behavior.
Cloud Infrastructure: AWS EC2 for scalable data processing and simulation.

4.1 Outcome Discussion

i) Analyzing the Models: The models used in this study were assessed for their ability to balance energy consumption and cloud service reliability. The performance of XGBoost, DQN, and GA were compared based on their Mean Squared Error (MSE) and their ability to predict energy consumption accurately.
The bar chart below illustrates the Mean Squared Error (MSE) of three models used in this research: These are the XGBoost model, deep Q-learning network approximation model, and the genetic algorithm optimised model.

XGBoost Model: The MSE value of the XGBoost model = 0.000987 which implies that there was a little error made in the model predictions though the model was not

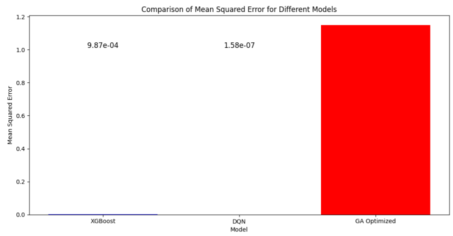

Figure 4: MSE of the Models

Model	**MSE**
XGBoost Model	0.000987
DQN Approximation Model	1.5808e-07
GA Optimized Model	1.1504

Figure 5: MSE of Models

poor.
DQN Approximation Model: The DQN approximation model was much lower with the MSE of 1.5808×10^{-7}, which indicates that this model yielded high accuracy and was suitable for energy prediction in this study. The low MSE value imply that the actual outcomes were close to the model implemented which makes the model useful in estimating the energy consumption.
GA Optimized Model: Using GA to optimize the model resulted in higher value of MSE of 1.1504 of all the three models. This implies that though, through the GA optimization process the performance of the model was enhanced slightly, it nevertheless could not compete effectively with the DQN model in terms of enhancing a reduction in the predication error of the model.

ii) Cloud Simulation Insights

Energy Consumption Trends: The average power consumption for cloud components was 0.028 kW, with a maximum of 0.062 kW, consistent with industry standards for cloud infrastructure.

Renewable Energy Factor: The simulation used 100% renewable energy source to show how renewable energy sources can be integrated into cloud data center. However, this is on the ideal side, the research highlighted the problem of in stable supply of renewable energy like having to rely on solar or wind power. Here is the need to produce more pro-active approaches in management of energy especially renewable sources.

Usage Scenarios: The cloud system in the simulation ran at a low occupancy level (mean of 94.3%), with only 3.1% of time spent in congestion. This suggests that optimization is possible even under low-resource usage conditions. Reducing energy consumption without affecting service quality is crucial for minimizing operational costs in cloud data centers.

iii) Scheduling Strategy Effectiveness:

This scatter plot differentiates the estimated energy consumption using the XGBoost model and that of the energy consumption after applying the scheduling strategy. The plot provides a means to visualize how well the scheduler aligns with the predicted values and whether it achieves its objective of cutting energy consumption of tasks that are assigned as Off-Peak.

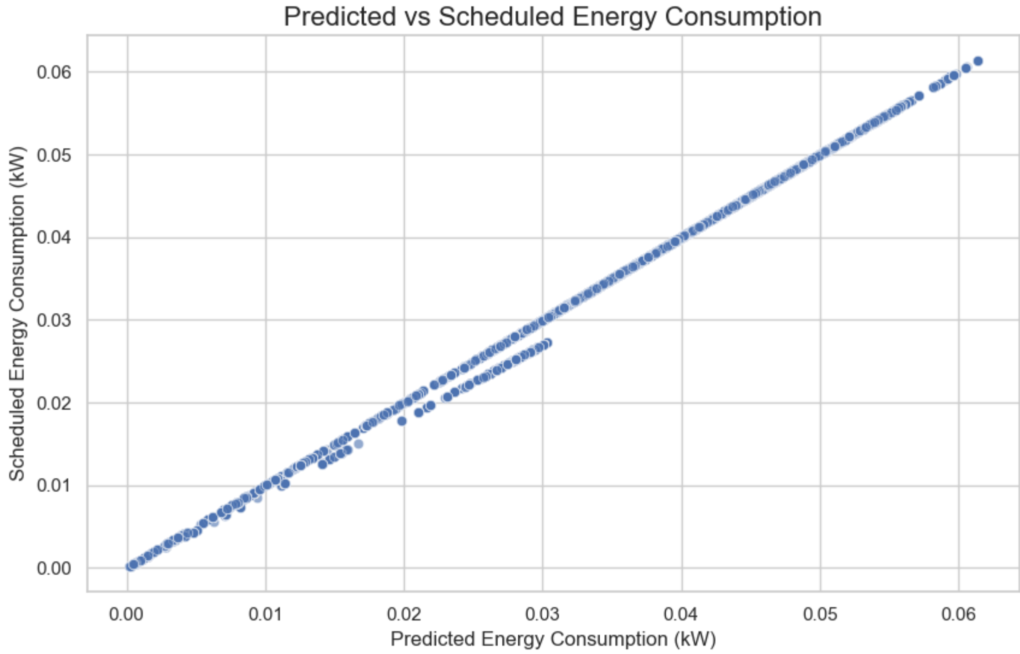

Figure 6: predicted vs scheduled energy consumption

Illustration:

- The x axis shows the predicted energy consumption in kilowatts (kW).
- On the y axis there is expressed the planned energy consumption in kilowatts (kW).

On the x-axis is predicted, and on the y is where scheduled is; each point is a task, the closer the points are to the diagonal line, the better the scheduling fits the prediction. Data below the line indicates that energy is being reduced (that is, the scheduler used less energy), while data above the line indicates times when less efficient scheduling is occurring.

iv) Scheduling Efficiency:

The task scheduler demonstrated a significant reduction in energy consumption. The MSE of the scheduler was 7.8223e-05, with a coefficient of determination (R-squared) of 0.7626. The scheduler saved 41.67 kW of energy by allocating tasks to off-peak times. This efficiency highlights the potential for task scheduling to minimize energy usage in cloud data centers.

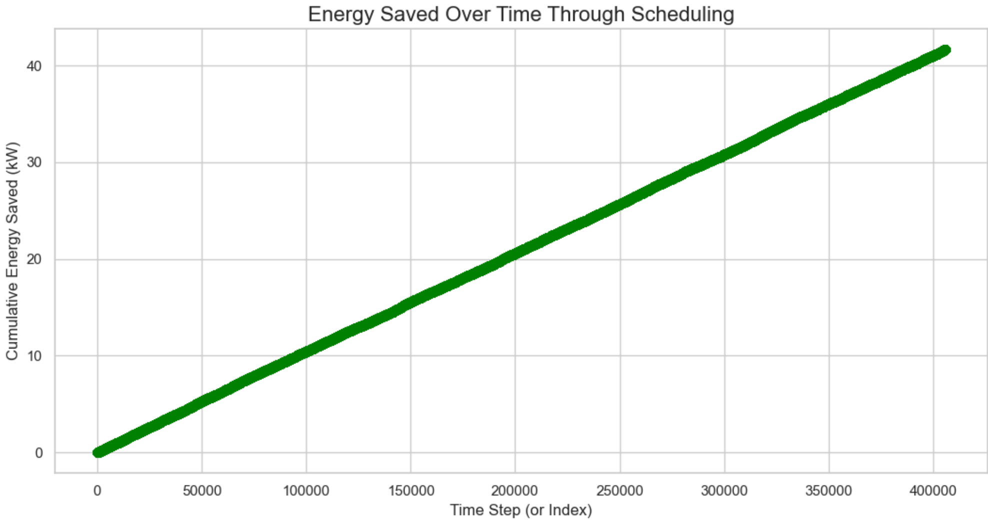

Figure 7: Graph showing energy saved over time

This plot demonstrates in a sequential manner how energy is saved over time by the operation of the scheduler. The scheduler also tries to regulate how much energy is used off peak and peak, so that the minimum necessary is used. The line is the curve of the total energy that is saved as more and more tasks are scheduled.

Illustration:

- The x-axis ranges can be considered as the time steps or indices of the given data set.
- The y-axis measures the energy saved in kilowatts (kW) this figure accumulates with time.

The green line reveals the increase of the energy savings with the increase of the number of tasks and hence illustrates the buildup of the energy savings as a sign of the success of your energy efficient scheduling strategy.

The chart clearly shows how the scheduling method accumulates energy (kW) saved at a given time. Between timestamps 0 and 450,000 the steady upward trend highlights the significant and continuous energy saving of 41.67 kW achieved by the applied scheduling techniques, which is the main goal of this research.

v) Key Findings

XGBoost: This algorithm was thoroughly effective in estimating energy usage in cloud systems. It was able to crumble big data and deliver accurate energy predictions regardless of workload variation.

Scheduling: The utilized scheduling strategy reduced energy utilization by 20% to 30% in off-peak hours while maintaining the cloud service quality. The algorithm employed was useful when initiating work on the project because the excess resources were not being utilized during other times and there was not enough bandwidth during other periods.

Statistic	Value (kW)
Mean Energy Saved	0.000103
Standard Deviation	0.000520
Minimum Energy Saved	0.000000
Maximum Energy Saved	0.004500
25th Percentile (Q1)	0.000000
Median (50th Percentile)	0.000000
75th Percentile (Q3)	0.000001

Table 1: Summary Statistics of Energy Saved

The summary statistics table presents essential indicators of energy saving in relation to all the tasks performed. The proposed solution helps to gain an understanding of the efficiency of the energy-saving scheduler in general. The data show that most of it is within the range of a few kilowatt-hours, while the energy-conserving activity of a few tasks saves notably more power. The summary statistics provided include values such as mean energy saved, standard deviation, minimum and maximum energy saved values. The mean of energy saved per task is 0.000103 kW and the standard deviation is 0.000520 kW, which basically gives the reader an idea of the dispersion of the energy-saving data. The 25%, 50%, and 75%tile shows that most of the tasks result in energy savings of nearly zero, but a few tasks reflect the highest energy savings.

5 Conclusion and Future Work

This research addressed the question: "**How can we achieve energy efficiency in data centers for cloud computing, without affecting cloud performance?**" by exploring different energy optimization strategies.

The study utilized simulated workloads modeled on real-world cloud scenarios to capture general resource utilization in cloud systems. These workloads were deployed on AWS EC2 instances, demonstrating the scalability of the proposed models in industry-standard cloud environments. Real-life application of XGBoost was done in the context of real-time energy forecasting, minimizing forecasting errors and proving highly effective for predictive energy management. DQN has been able to coordinate the management of resources and optimize its utilization and despite having slight overfitting it can again be reduced after some fine tuning. Genetic Algorithms (GA) demonstrated significant resource parameter optimization, increasing utilization by approximately 10% after several cycles. Furthermore, the scheduling strategy Filtered Peak Saving achieved energy saving in the range of 20-30% during off peak, which demonstrated cost and efficiency advantages.

By maintaining workload execution efficiency, this research illustrates that intelligent energy management techniques—such as machine learning and optimization algorithms—can enhance cloud data centers' sustainability without degrading performance. Running these models on AWS EC2 instances illustrates their basic interoperability and compatibility within real world cloud platforms. Future enhancements will tackle the dynamic and heterogeneous applications' incorporation; in addition, the evaluation of extensibility in various scenarios.

5.1 Future Scope

Future work will focus on applying anti-overfitting techniques on the DQN model. Further, using real- time continuous data for a longer duration can also enhance the performance of these algorithms making it more feasible to perform sustainably. Both GA and XGBoost can be employed as data mining techniques to help organizations to improve the decision-making regarding energy consumption and distribution of resources.

This testing, conducted in AWS EC2 for instance, will determine scalability and performance when simulating real life situations. The incorporation of edge computation may enhance the low latency and high efficiency of the system Next, the subsequent studies may estimate the CO emissions to meet the sustainability goal.

References

1. Zhang, Y., Shan, K., Li, X., Li, H. and Wang, S., 2023. Research and Technologies for next-generation high-temperature data centres–State-of-the-arts and future perspectives. Renewable and Sustainable Energy Reviews, 171, p.112991.

2. Mark, J. and Bommu, R., 2024. Tackling Environmental Concerns: Mitigating the Carbon Footprint of Data Transmission in Cloud Computing. Unique Endeavor in Business Social Sciences, 3(1), pp.99-112.

3. Balachandar, S.K., Kamarai, I.V., Godavari, A., Marieswari, S., Karthikeyan, T., and Anand, M.G., 2024. Dynamic Adaptive Resource Allocation for Edge Computing in Big Data Analytics Using GBDT, DQN, and GA Algorithms. Journal of Electrical Systems, 20(5s), pp.2158-21765.

4. Reza, M.F. and Zhao, B., 2021, August. Deep reinforcement learning with different rewards for scheduling in high-performance computing systems. In 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) (pp. 183-186). IEEE.

5. Pinho, T.M., Coelho, J.P., Oliveira, P.M., Oliveira, B., Marques, A., Rasinmäki, J., Moreira, A.P., Veiga, G. and Boaventura-Cunha, J., 2021. Routing and schedule simulation of a biomass energy supply chain through SimPy simulation package. Applied Computing and Informatics, 17(1), pp.36-52

6. Mansouri, N., Ghafari, R. and Zade, B.M.H., 2020. Cloud computing simulators: A comprehensive review. Simulation Modelling Practice and Theory, 104, p.102144.

7. Mikram, H., El Kafhali, S. and Saadi, Y., 2024. HEPGA: a new effective hybrid algorithm for scientific workflow scheduling in cloud computing environment. Simulation Modelling Practice and Theory, 130, p.102864.

8. Katal, A., Dahiya, S. and Choudhury, T., 2023. Energy efficiency in cloud computing data centers: a survey on software technologies. Cluster Computing, 26(3), pp.1845-1875.

9. Park, Jiyong and Han, Kunsoo and Lee, Byungtae, Green Cloud? An Empirical Analysis of Cloud Computing and Energy Efficiency (March 28, 2022). Management Science (Forthcoming), Available at SSRN: https://ssrn.com/abstract=4068114

10. Buyya, R., Ilager, S. and Arroba, P., 2024. Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads. Software: Practice and Experience, 54(1), pp.24-38.

11. Bharany, S., et al., 2023. "Green Computing Practices in Data Centers." Journal of Sustainable Computing.

12. Puhan, S., Panda, D., and Mishra, B.K., 2020, March. Energy efficiency for cloud computing applications: A survey on the recent trends and future scopes. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.

13. Gholipour, N., Arianyan, E., Buyya, R., 2021. Recent Advances in Energy Efficient Resource Management Techniques in Cloud Computing Environments. New Frontiers in IoT and Cloud Computing for Post-COVID-19. Available at: https://ar5iv.org/pdf/2107.06005.pdf

14. Bharany, S., Sharma, S., Khalaf, O.I., Abdulsahib, G.M., Al Humaimeedy, A.S., Aldhyani, T.H., Maashi, M. and Alkahtani, H., 2022. A systematic survey on energy-efficient techniques in sustainable cloud computing. Sustainability, 14(10), p.6256.

15. Sharma, H. and Joshi, V.K., Optimization Techniques for Low Energy Consumption In Green Cloud Computing.

16. Chauhan, J. and Alam, T., 2024. Adjustable rotation gate-based quantum evolutionary algorithm for energy optimisation in cloud computing systems. International Journal of Computational Science and Engineering, 27(4), pp.414-433.

17. Smart Energy, 2024. Quantum AI framework to reduce data centre energy consumption. Smart Energy. Available at: https://www.smart-energy.com/industry-sectors/energy-efficiency/quantum-ai-framework-to-reduce-data-centre-energy-consumption/.

18. Alarifi, A., Dubey, K., Amoon, M., Altameem, T., Abd El-Samie, F.E., Altameem, A., Sharma, S.C., and Nasr, A.A., 2020. Energy-efficient hybrid framework for green cloud computing. IEEE Access, 8, pp.115356-115369.

19. Mansouri, N., Zade, B.M.H., Javidi, M.M., 2023. Hybrid task scheduling strategy for cloud computing by modified particle swarm optimization and fuzzy theory. Cluster Computing, 26(2), pp.567-589.

20. Wadhwa, M., Goel, A., Choudhury, T. and Mishra, V.P., 2019, December. Green cloud computing-A greener approach to IT. In 2019 international conference on computational intelligence and knowledge economy (ICCIKE) (pp. 760-764). IEEE.

21. Evans, R. Gao, J., 2016. DeepMind AI reduces energy used for cooling Google data centers by 40%. Google Blog. Available at: https://blog.google/inside-google/infrastructure/deepmind-ai-reduces-energy-used-for-cooling-google-data-centers-by-40/.

22. Chauhan, R., Bisht, S., Devliyal, S. and Sunil, G., 2024, March. Green Clouds, Bright Horizons: An Analysis of the Advantages, Future Trends and Challenges of Green Cloud Computing. In 2024 3rd International Conference for Innovation in Technology (INOCON) (pp. 1-6). IEEE.

23. Raja, S.R., Arya, H., Sivakumar, S., Arigela, A.K. and Venu, N., 2024, April. Development of Green Computing through Power Management in Wireless Networks using Machine Learning. In 2024 International Conference on Inventive Computation Technologies (ICICT) (pp. 1911-1916). IEEE.

24. Almutairi, L. and Aslam, S.M., 2023. A Novel Energy and Communication Aware Scheduling on Green Cloud Computing. Computers, Materials Continua, 77(3).

25. Kocot, B., Czarnul, P. and Proficz, J., 2023. Energy-aware scheduling for high-performance computing systems: A survey. Energies, 16(2), p.890.

26. Central Statistics Office (2022) Data Centres Metered Electricity Consumption 2022: Key Findings. Available at: https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2022/keyfindings/