iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84790

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

Artificial Intelligence for Resource Optimization in Cloud Data Centers: Exploring Sustainability Challenges

Srikanth Gudi
Individual Researcher, Senior Software Engineer, 1251 Parkington Ave, Sunnyvale, CA, 94087

Abstract: There are concerns over cloud computing infrastructure as whether it could be sustainable for exponential growth of AI workloads. In this era, the present research examines the trends on energy consumption on cloud data centers and proposes that the use of energy efficient modules can help reduce the operational energy usage by approximately 70% without compromising the quality of service. This study suggests a holistic approach to green computing in the cloud for AI applications, which includes smart resource management, optimization of hardware and the use of modern cooling technologies. Based on industry facts, empirical evidence and key cloud provider reports, this study examines the priority issues in the current infrastructure and suggests solutions to these issues on the basis of empirical evidence. It is uncovered in the findings that more than 15% of energy is used by AI data centers and they are projected to consume up to 50% by 2030. It is clear there is a high level of consumption and fast action is required. This study evaluates the benefits of dynamic allocation of resources, virtualization, specialized AI accelerators, energy-aware scheduling and adoption of green energy. The study results indicate that the results of using several techniques in combination along with integrated techniques yield excellent results compared to interferences separately. The performance of a TPU based system is better than a general-purpose GPU, and AI-based cooling technology can reduce electricity consumption by approximately 40%. In addition, from on-premise to optimized cloud structure, migration can minimize carbon emission by 80%. This research contributes valuable, concrete and theoretical strategies for sustainable application of AI in cloud data centers, fulfilling a growing demand for computing.
Keywords: Resource Allocation, Cloud Data Centers, Carbon Emission, AI Models, AI Data Centers.

I. INTRODUCTION

But the most unexpected change has been the digital world on a global level – what Cloud Computing has brought about in an unexpected 'revolution'. Money that is for the environment – if it's going to cost, it's going to cost that's for sure. All global data centre energy consumption is 530 terawatt hours (TWh) which is less than 3% of carbon emissions. This will by 2030, be ~ 1000 TWh per year, which is the same as the Japanese's annual consumption. (Srivastava and Meenu, 2025; Deshmukh et al, 2025; Bolón-Canedo et al, 2024). The two technologies, the deep learning (DL) and the large language models (LLMs) have been fine-tuned and embraced by AI technologies in a drastic way over the last few years. Up to 15% of the total amount of power consumed in data centers may be consumed by AI workloads (Gabbatiss, 2025), which could increase by 50% by 2030.

The biggest carbon footprint is NOT related to the energy utilisation. The amount of Co2 emitted from one AI model (such as 626,000+ pounds per person AI model, compared to 5 times more that of an average vehicle). Also it will require special storage hardware for supporting AI hardware – which will have an impact on carbon footprints – and a massive cooling system for long storage – something that will take up a considerable amount of space. Otherwise, the digital transformation programmes could be unachievable and the amount of international climate targets through the cloud would be in jeopardy. There is a number of innovations however that can be utilized. But in the field of developing and improving green computing it's possible (it does seem!) to save a lot of energy without compromising performance.

The astute optimisation has led to a 40% saving in energy consumption in the cooling process, optimized by the Google's DeepMind team. Amazon Web Services (AWS) was able to bring down their consumption by 40% on Graviton architecture as compared to the traditional architectures (Aggarwal, 2025).

The results indicate that it is not necessary to compromise when it comes to the performance in order to be sustainable and computation can help with this. The principles of the Green Computing are: Energy efficient techniques, Energy from renewable resources, Smart resource management and Sustainable design.

It's an algorithm that makes use of hardware to the max, uses renewable energy resources, reduces the down time of resources and brings smart cooling into the cloud computing world.

This is because when scaled up, and made cloud based, the amount of energy infrastructure could be slashed by up to 50% and CO2 emissions by up to 88%.This shows its ability to have its cloud infrastructure transformed by 50% and to reduce its carbon emission as well by 88% (Srivastava and Meenu, 2025).

The clouds within an organization could be one of: Private Cloud, Public Cloud and Hybrid Cloud. This will prove to be useful when trying to implement cloud computing (Katal et al, 2023), and deciding on the platform that would be best suited for the enterprise. The new challenges in the energy consuming industry, combined with the evolution of the 'cloud' have resulted in an array of new technologies. The rapid growth of applications that are vital to certain businesses, including data storage, data processing and web hosting, is also being aided by cloud data centers. The enormous amount of cloud services puts the energy consumption of the data centres at more than 1% of the world's energy usage (Puso et al, 2023). This is an ideal situation from energy vendors and researchers' point of view, to help overcome a sustainability worldwide challenge that comes from cloud usage of energy.

With clouds in the picture, there are ample advantages with AI and ML to help optimize power consumption. While there are many benefits to be gained from the use of AI and ML in the cloud, one place where they can be a game-changer is power optimisation (Khapra, 2022).

These technologies can help use these resources in ways that won't waste them; and can offer potential energy savings that can be gleaned from the data. Intelligent resource allocation, adaptive workload scheduling, optimized power usage and optimized energy usage (including optimized cooling) of workloads are all possible with ML models. Leveraging its ability to analyse data on massive scales and conduct real-time analysis, AI programs can assist them in optimizing their cloud usage and make well-informed decisions to boost cloud efficiency (Gong et al, 2024). With good resource management, it's possible to dynamically use resources, which is a key application in energy, using ML/AI.

The traditional solutions are often an unwanted loss of energy, and are associated with over provisioning/underutilization of the resources (simple balancing, static pre-configuration). But for the other business, the ML models can predict what resources were used in the past which will eventually lead to flexible and accurate resource allocation, such as reinforcement learning, neural networks etc. (Fan et al, 2022) In hardware-based systems, the use of the ML models can be used to accurately predict failures of hardware equipment which can lead to reduction in waste, proactive maintenance and reliability of the systems (Kabir et al, 2022).

Regarding all of them there are important aspects to be taken into consideration: Smart load balancing, Energy conscious preparation.

AI can be used to schedule material and equipment usage during hours other than peak times to ensure that material and equipment are not unnecessarily used and minimise carbon emissions. Please be aware that this is especially in or on traditional resource dependent (TRD) areas (Zhang et al, 2018). However, the latter is more important and can support in terms of energy saving – optimisation of the cooling process. It is the data centres that consume a lot of power as they are the ones generating heat. Forecast the temperature change through AI which helps to control ON/OFF the Refrigeration/Air conditioning, to save power (Shukla et al, 2025).

For most companies that have sensitive data that is crucial, having data stored and protected in a private cloud network could be the solution that a big multinational company can guarantee, allowing them to take advantage of the cloud network. But, for the startup with limited budgets there are some of the less costly and more scalable services offered by the public cloud. The hybrid flexibility and security, however, is provided by the hybrid cloud solution as the organizations are now able to deploy critical data to the public cloud, whilst maintaining data security. Models might be regarded as a process of decision making which would be necessary in the future environment of the organization depending on its goals. The challenges such as increased energy consumption of the cloud data centers, and so on are significant to take advantage of the cloud computing capabilities of ML and AI (Chen et al, 2020). The cloud architecture in most cases is identical to the architecture outlined in figure 1, but the number of servers, as well as the architecture of each server are different.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

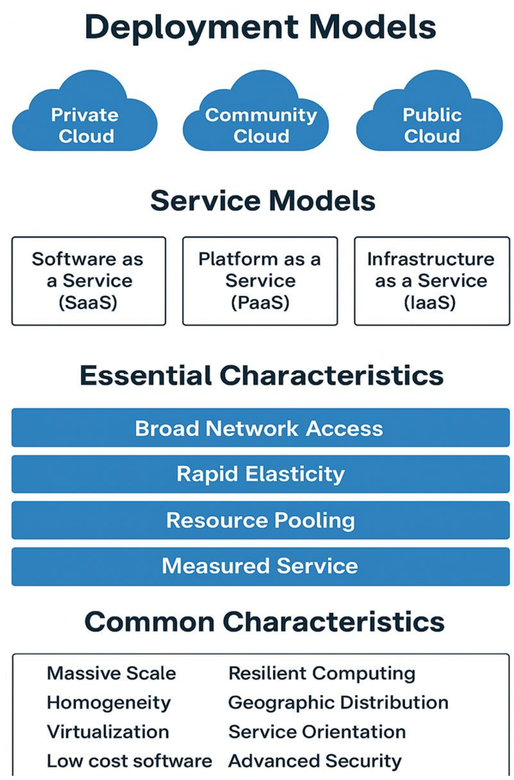


Figure 1 – A Flowchart of Cloud Infrastructure
Source – Chen et al (2020)

This might sound like a fantastic idea but there are several issues that need to be addressed as far as energy efficiency in cloud data centres and particularly AI/ML models are concerned. Not only do these solutions need to be scalableA and yet flexible when deployed in a dynamic environment supporting different workloads, they also need to be flexible in terms of deployment architectures (data centres and cross environment). The challenge is more complex when implementing such solutions in a dynamic environment in a large scale with AI models being flexible of different architectures (data centres, varying workload and multiple environments) (Khapra, 2024). Furthermore, the rising number of these applications is also accelerating (Duan et al, 2017), with many of these applications shifting towards privacy and security issues that can be addressed and resolved using AI. This will be followed by a couple of use cases, challenges and limitations that the researchers will be showcasing with the optimization of cloud data centres using AI and ML.

II. LITERATURE REVIEW

They are also widely utilized in the cloud resource allocation (Vakilinia et al., 2016) where the resources in the infrastructure are easily allocated and some simple heuristic algorithms, like deterministic Best-Fit ones and First-Fit ones (depends on local conditions) can be utilized. The baselines are simple, quick-to-use and rule oriented (static) scripts that are saved. They are not flexible structurally to make changes which occur dynamically. To address these issues, some additional heuristics based on evolutionary and nature, for example, "Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO)" have been applied to these, optimizing the use of the hardware and consumption of energy (Mireslami et al, 2019). While they may be efficient if they need to perform a certain number of calculations, if they are deployed, then realistically they will not be 'scalable' in terms of the number of people who can use them out-of-the-box as models are deployed and new ones added in real time.

One of the models that can lead to trapping at the initial stage of local optimum, is models which involve sudden changes in workloads which lead to fixed configuration. There are some of the natures used in recent years in meta-heuristics that are also applicable to static constraints such as "Whale" and "Spider-wasp" that can be used for finding optimal path, but all have their advantages and disadvantages in determining temporal anomaly one after another in streams of traffic. A model of "metaheuristic" that regulates tasks to shared uses of IoT has been developed recently. In contrast, the energy models are mainly focused on the local application of energy and not on multi-level protection against security attacks (Mohamed et al, 2025).

Srivastava 2025 first introduced the idea of integrated framework for green-computing for the cloud-based AI network which involved in the smart utilization of energy, green energy utilization, efficient utilization of hardware and advanced cooling technologies. They identified the key inefficiencies in the industry and, using case studies, a systematic review of the industry information and empirical evidence on some of the leading cloud vendors, evidence-based solutions were recommended. The study revealed that energy use in AI is over 15% of the energy that data centers use, and could reach up to 50% of the energy use by 2030. They claim to have a system with their TPUs that they wish to be able to outperform traditional GPUs in terms of performance/Watt. AI can also be used in optimized cooling, thereby saving 40% of the energy. Furthermore, carbon footprint can also be minimized by as much as 88% with transition to optimized cloud data centres.

The authors Xaba et al (2025) discussed about the allocation of resources in Virtual machines (VM) in cloud data centers which have the purpose of giving performance and scalability in the cloud networks. With increased demands and complexity of cloud environments, adaptive and efficient placement of VM is needed to retain optimal performance of networks. They have suggested a new modern optimization model – NetDEO which allocates the resources according to real-time real demands by PSO. In another aspect, when looking at major topics such as traffic stress, load balancing, and the smooth workload adjustment in different networks, NetDEO drastically reduces traffic, balances load equally and optimizes the placement of VM. But, the performance of simulation NetDEO was better than the conventional methods.

To accomplish both of these goals, Al Masarweh et al introduced two frameworks “AI-driven Monitoring (AIDAM)” and “Geographically Aware Placement Algorithm (GAPA)”. For optimizing workload share according to the capacity of local server and proximity of the servers, GAPA was used and for the threat reduction and real-time anomaly detection, deep encoders were used by AIDAM. The framework was tested by using the deterministic simulation, for the corpus “Google Cluster Data (2019)” and in the presence of DDoS anomalies. It's been found that rather than physically re-engineering infrastructure, they can develop a more resilient and secure secured infrastructure.

III. RECENT DEVELOPMENTS IN AI-BASED RESOURCE ALLOCATION

A multi-layered and systematic approach is needed to save energy in a cloud computing system as all the components of an operational, software, hardware and an infrastructural system should be taken care of. The framework follows a few fundamental concepts which will be used to describe various activities to leverage the green computing in various layers. The first is to think about how the system can be made more efficient, in terms of energy consumption, at each phase of the data center’s life cycle - starting from hardware selection, design, operation and software development. Secondly it should be optimized by “use to the max”, which will minimize the idle usage of the resource and maximize to get better performance. Thirdly, computation demands needed for integration of the green energy, on the supply-side and the demand-side. Dynamic changes on the other hand requires proper monitoring and optimization based on the “current” conditions, which should be in line with the changing environmental conditions and workloads (Wilson, 2024).

A. Resource Allocation Requirements.

1) Robust Metrics

Proper optimization and measuring the energy efficiency of various areas of cloud network requires the usage of measurement and metrics. Industry measure to enhance data center efficiency is “Power Usage Effectiveness (PUE)” which is ratio of Data center electricity use to the electricity used by the IT equipment. But there are some constraints on PUE as a full-fledged sustainability measure. It's a Zero that is embodied Carbon, Zero water use and Zero carbon intensity. Some new metrics plug the gaps – such as Carbon Usage Effectiveness (CUE) for carbon emissions per unit of IT equipment – and the Water Usage Effectiveness (WUE) analysis of the water consumed in cooling cloud data centers as well as in energy generation. Lavi (2025) says the effectiveness of the use of IT hardware has two metrics: the operational carbon intensity (grams per kWh of carbon equivalent) and the performance per watt of IT hardware.

If workloads are based on AI, it is essential to incorporate performance across the workload, as this will take into account the carbon footprint and computational needs of the AI workloads. So, to compare the different architectures and hardware platforms we can make calculations of computation per Watt. During training, the quality of the model can be evaluated along with efficiency in terms of watts of power, during inference when the model is applied, the latency/throughput in terms of watts of power. The use of the grid intensity is for actual carbon footprint of data centres (Cordingly et al, 2023).

2) Hardware Selection

Finally, the secret of cloud data centers' energy efficiency that are heavily reliant on AI is the type of hardware that is used. The amount of hardware used will be in direct proportion to the carbon emission impact and long-term cost in the future. In today's day and age, the data centre is where an array of technologies can be used to fuel a variety of workloads including GPUs, CPUs, AI accelerators and tensor units. The solution of relying solely on raw power is not enough, neither is the platform for some workloads – each of these can be a big determinant of system efficiency. For transformer networks, like those used in an LLM, scaling up training of such models and inference with neural networks using the TPUs is best suited. The GPUs are capable of a variety of scientific computing, machine learning and graphics processing applications. The new generation of the chips are "balanced" AI chips, the latter of which have AI acceleration capabilities towards hybrid workloads – workloads that involve the use of AI and traditional workloads (Mavridis and Karatza, 2019).

3) What model(s) and software(s) are the optimum ones for optimization?

In some cases, the energy consumption can be optimized to a higher level without changing the hardware, in other cases, however, it may be necessary to change the hardware – and in some cases an energy efficient optimisation that is cost-effective. It is cost-effective to optimize the energy use in the cloud infrastructure along its way to become more sustainable, as Agdestein (2025) aims to do. Computational needs and energy consumption can be affected drastically with selection of model architecture. Depending on the type of model, a task-centric model can be extremely energy efficient (e.g., 5-10 times less computation than dense model) when compared to the energy demands of large models and for specific applications; a limited number of models can save much energy (Yunus, 2025). Exploiting the right patterns (such as optimizing for attention mechanism, depth-wise challenges and activation functions etc.) can be helpful in reducing the inference and training costs significantly without affecting accuracy.

Optimization of training process is an approach that can minimise energy utilization in training which can be applied using several approaches. The length of training as well as total power consumption can be influenced by tuning the parameters such as choosing the batch size, scheduling the learning rate, and the gradient accumulation strategies (Yunus, 2025). If the efficiency value of the validation set (Y) does not change, then the training process will also be terminated, because if the training process is carried out, then it will not be able to improve the efficiency value of the validation set (Y).

4) Smart Resource Utilization

In a cloud data centre, two key points in being energy-efficient are: Scheduling of workload and Allocation of resource. They can be used to provide resources when and where needed, rather than provisioning over and over again (Svystun, 2024) and can be used to rapidly scale up and down resources based on need. Cloud can be used to scale-up/down the resources based on requirements. Based on workload metrics, the cases are removed and/or added to solve the problem of saving energy with not an excess use of energy. This is complemented by the reduced carbon footprint of serverless computing, since it only uses the resources when the functions they handle are used, which can further cut down the carbon footprint by 88% compared to AI workloads on data centres (Kräuter, 2024). AI powered resource optimisation (Oloruntoba, 2024) ML can forecast resource workload and proactively assign resources. Besides, predictive analytics could also be employed to forecast the peak of demand and scale up the resources ahead of time to attain the best performance while not consuming a lot of resources (Gopalaiah, 2025). Each month, 12 MT of carbon emissions are cut via the use of ML models to manage storage for Microsoft, and identify underused virtual machines (Pecan, 2024). The energy usage anomalies can be detected automatically and hence the anomalies are automatically notified at real-time. Efficient workload sharing (Svystun, 2024), this is achieved with the help of a number of applications for visualization of workload.

With the use of the servers on shared servers, a wide variety of utilizations can be achieved ranging from 10% utilizations to 60% utilizations (Mavridis and Karatza, 2019). This collaboration reduces the needs of servers, and consequently decreases energy consumption of the cooling and computing needs of an information centre. Efficient use of virtualization resources, virtualizing the OS kernels within containers (Srivastava and Meenu, 2025), separation of applications from virtualizing the OS kernels in different containers and virtualizing across containers with technologies like the Kubernetes and Docker which further reduce the load of virtualization.

B. Recent Developments

Many studies have been carried out on optimizing the energy consumption in a cloud data centre – using models based on ML and AI. The majority of these studies are based on various models with different themes: predictive maintenance models, resource management models, load scheduling and load balancing models, energy aware models, cooling optimisation models as well as renewable energy.

1) Resource Allocation

To conserve energy, and improve performance, it is very important to optimise resources ideally in cloud data centres. The studies for smart allocation of resources would be carried out based on the machine learning (ML) and artificial intelligence (AI). Furthermore, reinforcement learning can be leveraged to optimize various tenants' cloud resources with respect to various workloads and its energy consumption and performance (Gupta and Katiyar, 2016; Radhamani and Dalin, 2019). The Supervised Machine Learning models can predict the trend with a good accuracy which can be helpful in provisioning of the resources and hence energy and energy consumption can be saved (Rathor et al., 2015). The Cloud Data Centers (CDCs) and Hybrid provisioning and deep reinforcement learning (DRL) have been helping virtual machines (VMs) in the cloud (Shrimali and Patel, 2015) to save energy.

2) Maximum Scheduling & load balancing

They are optimally loaded on the servers and can be scheduled depending on the energy requirement of the server, to reduce the energy consumption. The dynamically load balancing based on machine learning model (Li et al., 2021): Dynamically balance the load according to information at the moment; reduce the energy consumption of the server; and avoid the server overload. In addition to analyzing real time data and past experience, an AI scheduling model can also predict what amount of resource and time a workload will take, based on some criteria, and reserve a resource for a workload. To minimise the operational costs of the acts and power consumption, deep reinforcement learning is used to optimise task scheduling by reducing the need of power consumption (Dabbagh et al, 2014).

3) Predictive Maintenance (PM)

Leveraging AI, PM can predict equipment failure and prevent energy loss and equipment down time due to failure. The sensor data driven approach can use the ML approach, using previous failure reports. It can be beneficial in preventing inefficiencies in energy use (Mughees et al, 2020) and it can assist in making proactive maintenance decisions, by considering the risk of hardware failure. Using the models developed with DL, the health of the system can be monitored, and events in the system can be predicted before it causes any unwanted outages thus reducing losses and increasing system reliability (Srilatha, 2019; Han et al, 2015). The AI Models identify the deviations from the energy saving energy pattern and optimisation time is determined based on the energy saving optimisation for maintenance. (Usman et al, 2019)

4) Energy-Aware Models

In the developed models, several research has been carried out on energy field of core operation that is energy saving. Because of new roles and activation scheduling of these roles considerable energy saving without compromising the service quality is achieved by these roles allocation's power saving is taken in to consideration. A resource management/scheduling based different optimization-based model has been proposed (Khan et al, 2019) which is based on minimizing the total energy consumption in cloud. The energy efficient models are fine-tuned using the assistance of ML and optimized models to match with the workload requirement.

5) Optimizing Cooling Systems

One of the largest amounts of energy consumed in a data centre is energy usage – and with efficient use of that energy, the use of ML and AI is possible. Research on cooling systems that use AI have primarily been concerned with how to design predictive maintenance to predict the temperature changes and modify the cooling design to cut energy usage without compromising performance (Das et al, 2018). Energy saving cooling options can be used to moderate thermodynamics; and the parameters can be set up to achieve an energy balance. The carbon factors and workloads will be optimized for best energy conservation and the RL methods will attempt to optimize the cooling during low demand time, effectively attempting to tune the cooling towards optimisation.

6) To take advantage of Renewable energy resources with the support of AI

Another area of research that is important is to enhance the use of green energy in cloud infrastructure. Some studies related to integration of renewable energy with artificial intelligence while some other studies using predictive model for the estimation of wind and solar energy resource availability for re-scheduling and allocating them to reduce dependency on fossil fuel energy resources (Kaur et al., 2020). Energy consumption in the scheduling optimisation of the workloads in sustainable cloud data centres is huge.

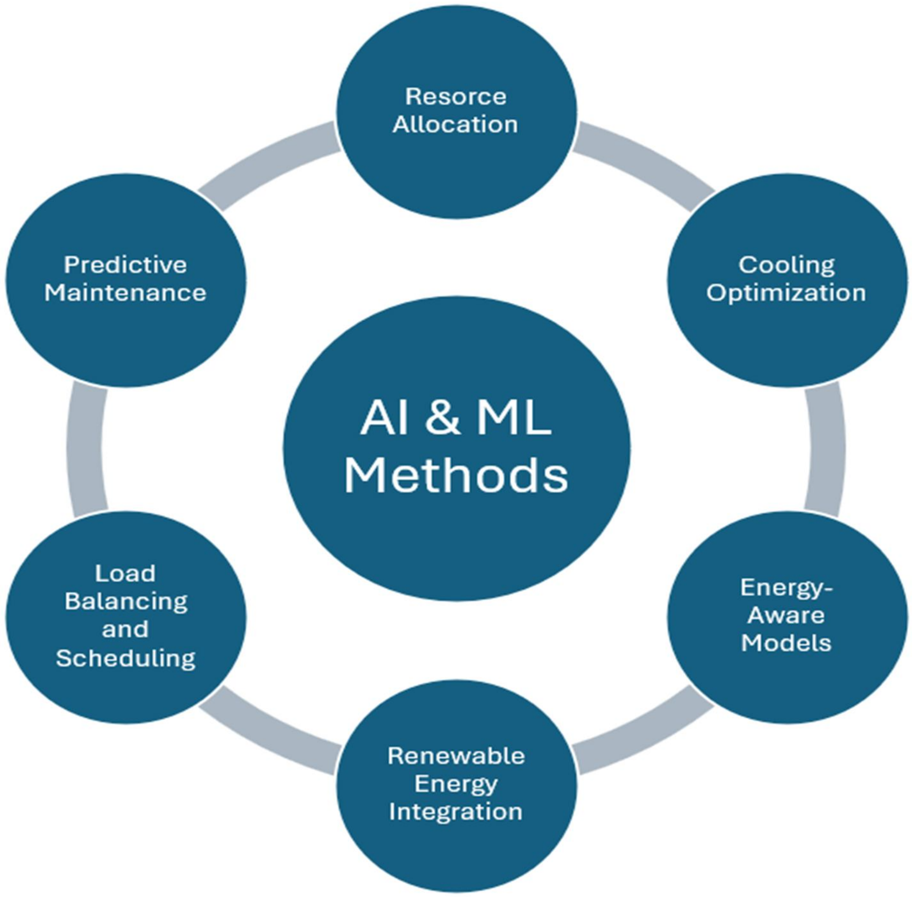


Figure 2 – AI and ML based Approaches for Resource Allocation

Figure 2 illustrates approaches based on AI and ML in VM environments, such as, load balancing, allocation of resources, cooling improvement, predictive methods, renewable energy integration, and energy-aware models.

IV. FILLING SUSTAINABILITY GAPS WITH CARBON-AWARE COMPUTING

The importance of computing in significant GHG emissions (2,5% - 3,7% of the global GHGs) (Ferreboeuf, 2019) puts it on the agenda for organizations concerned about reducing carbon footprint and makes it an important topic of discussion and conversation. The re-Engineering of the software can also be carbon-smart and adopting hardware solutions can be implemented to extend the product life and/or redesign the data centres. The carbon-aware computing is a relatively new computer software sector – computer software capable of knowing the carbon emissions – employs computing resources to adjust their utilization in accordance to the power grid. The concept of 'carbon aware computing' is the more computing, the more carbon the power supply uses; the less computing, the less carbon the power supply uses. The above can be achieved by employing – Optimal location of server in the area of the Grid. Choose the position of the server in the "green" space of grid. For the processing jobs which are not sensitive to the running time such as training of ML model, batch processing jobs with maximum % of green energy in the grid. Shaping demand – Software will only be allowed to be running, but at a lower emission rate – at high emission times. But if these are not something that the infrastructure vendor can easily do, then it's time for organizations to come up to the plate! For others that don't have these abilities, they can simply update their software to make it carbon-aware, and reduce emissions greatly without having to make their software carbon-aware. There are some differences in composition of the grids between the best and the worst but the changes in the grids on a day-to-day or month-to-month basis are not as great in region.

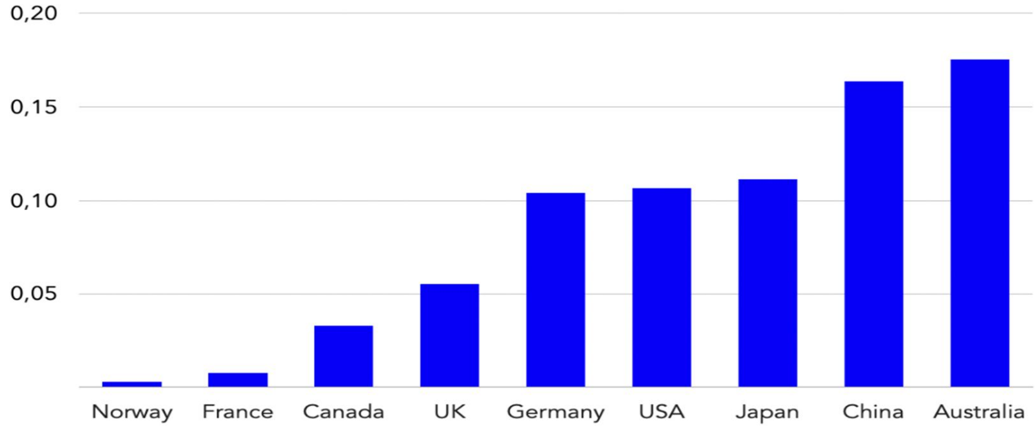


Figure 3 – Region-wide variations in carbon emissions (kg CO_2e per kWh) since 2021
Source – Wengel (2025)

The most important factor when minimizing the carbon footprint of a cloud data centre is to select an area that has a 'highly green grid' that the software can be installed on. It should be decided where the Cloud Data Centers will be located in terms of carbon emissions - 'best'. The carbon emissions from the local grid blend that is used are variable over time, since the proportion of different energy sources in the grid and energy demand fluctuate over time, and because the proportion of energy produced from renewables on the grid (renewables content) also fluctuates hourly, as with the US. In some of the squares there are some patterns – During daytime, more renewable energy could be generated due to solar energy in the grid. More electricity usage at home typically is on the afternoon/evening of the day.

More often, it's more practical to generate the green energy at night, when there is a lot of wind energy.

Although it only affects the amount of carbon used in powering servers, scheduling usage and proper deployment of these servers could lead to a decrease in carbon emissions at cloud data centres. The lion's share of emissions from cloud data centers is also from personified emissions, in addition to the usage of energy. Disposal of server, personified emissions: emissions from server and components. The electricity use and use of the consumer electronics accounts for 75-85% of the emissions, while the remaining 15-25% of the total emissions. Likewise, carbon-aware computing doesn't save energy throughout the PC's life cycle, just from the server; when the PC is discarded and when it is manufactured.Similarly, using carbon aware computing, it only saves part of the electricity that the server consumes, not the life of the server – either that you use, or that the server uses in your production. The use of the hardware should also be optimised, prior to replacing the same (Circular Computing, 2021) and optimisation of the load on the servers should also be optimised. Normally, the servers replacement period of the cloud data center is 4-5 years. So they have to be utilized to the maximum extent possible. Reducing the number of physical servers to around 100% utilization and the use of rented servers, will help to reduce carbon footprint. Moreover, it can also maximize the resource utilization, and lower the working costs (Wengel 2025).

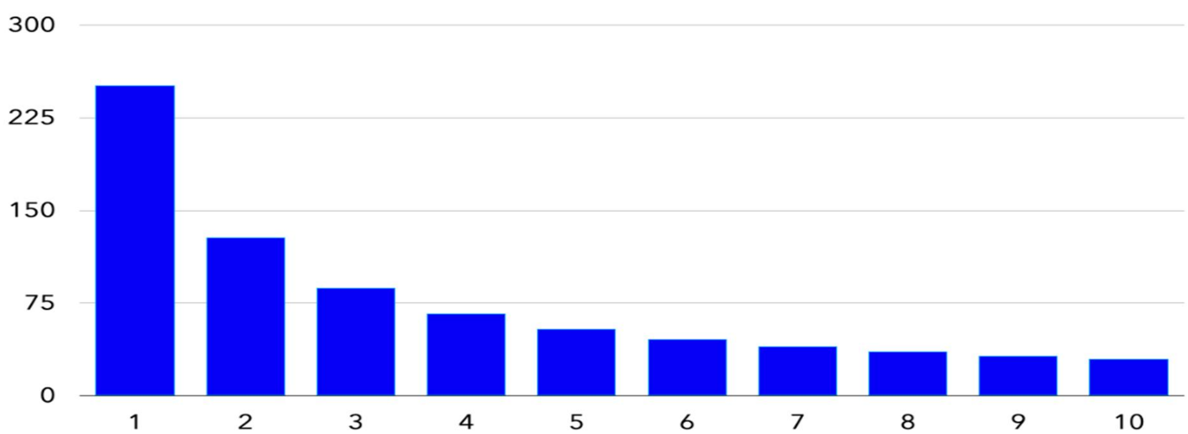


Figure 4 – Carbon emissions per job (kg CO_2e), i.e., 1 job needs 10% of capacity of the CPUs
Source – Wengel (2025)

According to the article there are also ways of optimizing the cost of carbon emission of computing, in addition to infrastructure to run other services.

There is a need to avoid spending time excessively on building software that reacts to power grid dynamically as there are marginal benefits.

The cloud regions have to be comprehended along with the cloud infrastructures and in a sustainable way (with regard to the grid). This is normally offered as a part of the cloud service providers.

Further studies should be undertaken to explore the potential to change load and location and decrease carbon emissions from cloud service providers.

That's where the Cloud providers could benefit from the underutilized instances, using greener instances, such as Amazon Web Services' spot instances to do our jobs – and save costs and carbon.

Just like the cloud providers would use their VMs, which they are renting, to their maximum capacity (as managed or serverless VMs) to make best use out of the provisioned infrastructure, the same applies to the VMs rented as well.

A. *Carbon-Aware Initiatives*

Being energy efficient and environmentally responsible are increasingly becoming valuable assets in businesses (Bioscore, 2024) and cloud service companies are at the forefront when it comes to using green energy to minimize their carbon footprint. Google Cloud was the first big cloud vendor to become 100% green energy in 2020, with a purchase of energy credits and contracts for direct energy purchases (Budinski, 2023). The company has installed more than 500 wind and solar energy projects around the world, and purchased 7.2 GW worth of green energy (Oberhaus, 2019). Google isn't only matching green energy, but looking to make 24x7 renewable energy in the distant future.

The aim is to make a concerted effort to become 100% green energy by 2025 (Hivenet 2025) and carbon negative by 2030. As reported, Microsoft are also signing new contracts for 72% of their consumption of green energy and they even go so far as to investing in green energy projects (Aggarwal, 2025). It is decarbonized and will sequester all carbon emitted since it was established in 2025 (Derrick, 25) and it has to remove a lot of carbon from the atmosphere technologically or naturally.

AWS (Bioscore, 2024) has the most implementation of green energy projects in the world compared to other corporate energy buyers, where they have over 380 global green energy projects. On the local level 100% green energy can be achieved with 19 AWS (achieved in 2022). But, until now, there have been concerns about how transparent AWS is with regards to its emissions – organisations have not been keen on releasing detailed data about individual AWS operations at its centres (Oberhaus, 2019). Even if a good procurement of green energy is achieved the important issues of impact and legitimacy of green energy commitments remain. Whatever the amount of green energy procured might be, there are still some doubts regarding the impact and reliability of green energy. The production of most of the green energy is actually of the “renewable energy credits (RECs)” variety, and not the direct green energy (Oberhaus, 2019). Although the Organisation are still reliant on energy from fossil fuel sources, through the use of RECs (Oberhaus, 2019) they can say that they have green energy. The data centre will then return to the other gas provider(s) to purchase gas credits, and now has as many gas credits as they purchased gas. This practice, which contributes to green benefits without using green energy in the future has been criticized (Oberhaus, 2019).

B. *Carbon-Aware Scheduling*

Carbon aware computing has been a trend over the past few years, with the intent of improving the carbon intensity of the available energy by scheduling the computation to the right time and right place, when green energy (Ekanayake et al, 2023) can be made available. This can be achieved through a time frame by altering the carbon intensity of grid, or geographically by decreasing carbon emission without any extra green energy installations as discussed in Ekanayake et al (2023). By using the optimal location to train AI models, carbon emissions can be lowered by more than 1000% compared to other average locations (Ekanayake et al, 2023), while also using the right time in the training process to lower carbon emissions by 50% (Ekanayake et al, 2023). Carbon-aware directing in the cloud can be achieved by providing physical distribution of computing resources in several regions – that could be different energy supplies. (Jiang et al, 2024) Schedulers can be available around various areas and computing of routes to areas where greener energy is available (Jiang et al, 2024; Cordingly et al, 2023) can drastically affect carbon intensity. This will help to reduce many of the delay tolerant operations such as transcoding, batch processing, training of ML etc.'s carbon footprint without impacting the user experience. This flexibility of assigning resources at a very fine level of granularity, coupled with the flexibility of moving resources in terms of location is well suited to optimise the carbon-aware computing on these platforms (Cordingly et al., 2023).

The validation of effectiveness of the carbon aware scheduling is carried out by the empirical studies. The carbon intensity was shown to be reduced by over 99.8% for over 19 regions globally with the serverless platform, and to be reduced by 65% with optimisation goals (Cordingly et al, 2023). A maximum reduction of 10 times in carbon footprint has been obtained by scheduling the ML loads with the proper scheduling (compared to the greedy/random scheduling of the ML loads (Ekanayake et al., 2023)). This is achieved using geographical diversity in the form of energy mix – marginal carbon emissions at various locations in the grid and time varying green energy (Cordingly et al, 2023). Time-shifting, low-grid carbon footprint and flexible workload scheduling - one can adjust the workload to the times when green energy is produced in a large amount - could be yet another element for carbon-aware solutions added in the future. The wind and solar images will provide schedules, which can be forecasted, based on information on weather seasonal and daily patterns. A ML can be applied to predict usage of renewable energy (RE) resources and grid intensity, and schedule jobs on the basis of the prediction to maximize the use of the renewable energy. These delays are potentially reversible ("backwards"), and can be reduced by extending the AI models that perhaps have had to wait for these delays.

V. CHALLENGES AND FUTURE DIRECTIONS

A. Challenges in Optimizing Energy Usage

Irrespective of increasing commitments towards green energy, there are several economic and technical issues that obstruct shifting to cleaner energy for cloud data centers (Jiang et al, 2024). The sporadic nature of wind and solar energy generates variations in supply which should be managed over consistent energy demand of data centers. Generation of solar energy follows foreseeable cycles and variability which should be managed over constant energy demand by data centers (Budinski, 2023). Expected daily cycles are followed by solar energy generation without producing power at night, while wind energy generation depends upon weather which may not be aligned with increased demand of computation (Budinski, 2023).

This variability needs either backup of generation of fossil fuel, energy storage, or depending on grid energy from different sources when renewable energy is not generated efficiently (Oberhaus, 2019). While improving, technology related to energy storage has been insufficient for round-the-clock operations at the scale of data centers. Battery-operated storage systems can buffer changes in the short-term without having cost-efficiency and capacity for multi-day storage needed to manage longer periods of low power generation (Budinski, 2023). Alternative options are storing compressed air energy, pumped hydro-power storage, and hydrogen production. These are promising technologies but have location and infrastructure dependence (Budinski, 2023). Until the availability of long-term storage, either location in areas with consistent resources of renewable energy or fossil fuel are needed to achieve complete carbon-free operation round-the-clock. The geographic delivery of ideal resources is not aligned with optimal locations of data centers evaluated by other factors like availability of land, user proximity, regulatory issues, and water access (Bolón-Canedo et al, 2024). With excellent wind or solar energy availability, there are some locations where there is limited connectivity with fiber optic network, supply of cooling water, or enough space for deploying big data centers (Bolón-Canedo et al, 2024). On the other hand, areas with ideal characteristics and existing infrastructure for data center may have limited potential for green energy (Bolón-Canedo et al, 2024). This physical disparity adds conflict among location dependency and green energy goals. There are also other challenges at the scale of data center for integrating green energy. Hundreds of megawatts of energy are needed for large data centers, which can be provided with huge electrical infrastructure to work with power grids (Budinski, 2023). There is limited capacity for transmission of that much power in some areas. So, it becomes difficult to transmit green energy to data centers from generation sites (Budinski, 2023). There is planning, construction, and permission required for expansion of grid upgrades to support expansion of data center and green energy, which may add delays in deploying green energy.

B. Future Directions

There are pathways available from emerging approaches to achieve completely reliable and green cloud computing powered sustainably. Cutting-edge technologies used for energy storage are liquid energy storage, flow batteries, gravitational storage, and thermal storage for seasonable or multi-day backup at reasonable prices (Budinski, 2023). With electrolysis, hydrogen can be produced with excessive green energy to provide both clean fuel and energy storage for backup. Carbon-free power can be provided by small modular reactor (SMR) units for data centers to fulfill the needs with variable green sources. Direct generation of green energy at data centers provide benefits like reliability, low transmission losses, and simple accounting of renewable energy. On-site installations of rooftop solar panels on data centers and land can generate a lot of clean energy, though usually limited to meet complete demand of facility. Data centers also face environmental and zoning limitations for wind turbines at data centers but may be ideal in some areas (Dsouza, 2025). Grid energy can be combined with on-site power generation to deliver hybrid systems to maximize the use of clean energy.

Advanced approaches are demand response initiatives, virtual power plants, and bilateral agreements to acquire power for smarter use of clean energy. Those plants collect distributed resources like battery storage, backup generators in data centers, and flexible loads to improve integration and grid services (Dsouza, 2025). Data centers are incentivized by demand response programs to shift or minimize consumption during grid stress to stabilize grid and integrate renewable energy (Dsouza, 2025).
Direct transactions are enabled with bilateral agreements between green energy generators and data centers, offering certainty in revenue for green energy projects and securing computing. Though in nascent stage, quantum computing (QC) could minimize power consumption drastically for some computational issues. QC can boost classical models exponentially for certain roles like simulation, ML and optimization. If QC becomes scalable, energy consumption could be reduced by solving complex issues with limited resources than traditional systems. However, there are challenges related to QC like limited number of qubits, error rates, and very low temperatures to increase energy demand (Gopalaiah, 2025).

VI. CONCLUSION

This study includes full analysis on green computing for AI data centers. It is found that energy can be conserved to a large extent and carbon footprint can be minimized by incorporating smart energy management, hardware optimization, advanced cooling, software efficiency and integration of green energy. As per the evidence, specialized AI data centers can deliver better performance up to 2-4 times than general processors. By optimizing cooling using a artificial intelligence, energy can be saved by 40% and by migrating to the cloud, carbon emission can be lowered by 88%. The enhancements in efficiency offer huge advantages in the environment and systematic implementation offers a better computational output. But aggregate impact of cloud computing continues to be substantial and there's a substantial increase in overall demand for computation, particularly for AI applications. AI workloads are expected to make up 50% of data centers' electricity consumption by 2030, which will more than double their usage. This expansion can turn out to be a danger to the efficiency improvement. It must put in place sustainable practices as soon as possible. This urgency is increased by the issues of climate change. A sustained decarbonization is required for all sectors, such as IT. For sustainable cloud computing, technical foundations have been set up. To ease the burden on ML training, there are several AIs accelerators which require efficient platforms to support this. The power densities of the immersion and liquid cooling technologies are achieved at lower costs for energy production. Resource usage can be improved and idle energy consumption can be reduced by containing and virtualizing them. Cloud operators can take full advantage of power agreements and direct purchase of clean energy. Workload can be distributed as per grid intensity with carbon-aware scheduling.

REFERENCES

[1] Agdestein, I. (2025). AI Energy Efficiency: Reducing Power Consumption in AI Models. Focalx. Available at https://focalx.ai/ai/ai-energy-efficiency/
[2] Aggarwal, S. (2025). Cloud Sustainability: Green Initiatives and Energy Efficiency. Tech Ahead. Available at https://www.techaheadcorp.com/blog/cloud-sustainability-green-initiatives-and-energy-efficiency/
[3] Al Masarweh, M., Alwada'n, T., Hamdan, A. M., Almomani, O., & Mustafa, I. A. (2025). Optimizing Resource Allocation and Enhancing Security in Cloud Systems: A Data-Centric Approach. Computers, 15(7), 419.
[4] Bioscore (2024). Green Cloud Computing: A Comprehensive Comparison of Cloud Providers' Sustainability Efforts. Available at https://bioscore.com/blog/green-cloud-computing-a-comprehensive-comparison-of-cloud-providers-sustainability-efforts
[5] Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, 128096.
[6] Budinski, M. (2023). A Green Cloud Race: Microsoft, Google, and Amazon Compete for a Sustainable Future. Deployflow. Available at https://deployflow.co/blog/a-green-cloud-race-microsoft-google-and-amazon-compete-for-a-sustainable-future/
[7] Chen, Y., Liu, Z., Zhang, Y., Wu, Y., Chen, X., & Zhao, L. (2020). Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things. IEEE Transactions on Industrial Informatics, 17(7), 4925-4934.
[8] Circular Computing (2021). What Is The Carbon Footprint Of A Laptop?. Available at https://circularcomputing.com/news/carbon-footprint-laptop/
[9] Cordingly, R., Kaur, J., Dwivedi, D., & Lloyd, W. (2023, September). Towards serverless sky computing: An investigation on global workload distribution to mitigate carbon intensity, network latency, and cost. In 2023 IEEE International Conference on Cloud Engineering (IC2E) (pp. 59-69). IEEE.
[10] Dabbagh, M., Hamdaoui, B., Guizani, M., & Rayes, A. (2014, April). Energy-efficient cloud resource management. In 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 386-391). IEEE.
[11] Das, K., Das, S., Darji, R. K., & Mishra, A. (2018). Survey of Energy-Efficient Techniques for the Cloud-Integrated Sensor Network. Journal of Sensors, 2018(1), 1597089.
[12] Derrick, M. (2025). What is the Truth About Future AI & Data Centre Emissions? Data Centre Magazine. Available at https://datacentremagazine.com/news/what-is-the-truth-about-future-ai-data-centre-emissions
[13] Deshmukh, V. D., Singh, S., & Waghmare, S. (2025). Green Cloud Computing: Energy-Efficient Solutions for Sustainable IT Infrastructure. International Journal of Mobile and Cloud Systems Engineering.
[14] Dsouza, R. (2025). How Green Cloud Computing Drives Sustainable Business Growth in 2026. Bacancy. Available at https://www.bacancytechnology.com/blog/green-cloud-computing

[15] Duan, H., Chen, C., Min, G., & Wu, Y. (2017). Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems. Future Generation Computer Systems, 74, 142-150.

[16] Ekanayake, S., Shah, T. and Evans, S. (2023). Optimizing emissions for machine learning training [Presentation]. GE Vernova. https://www.gevernova.com/gev/sites/default/files/2024-07/presentation_optimizing-emissions-for-machine-learning-training.pdf

[17] Fan, G., Chen, X., Li, Z., Yu, H., & Zhang, Y. (2022). An energy-efficient dynamic scheduling method of deadline-constrained workflows in a cloud environment. IEEE Transactions on Network and Service Management, 20(3), 3089-3103.

[18] Ferreboeuf, H. (2019). Lean ICT: Towards digital sobriety. The Shift Project. Available at https://theshiftproject.org/en/publications/lean-ict/

[19] Gabbatiss, J. (2025). AI: Five charts that put data-centre energy use – and emissions – into context. Carbon Brief. Available at https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into-context/

[20] Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). Dynamic resource allocation for virtual machine migration optimization using machine learning. arXiv preprint arXiv:2403.13619.

[21] Gopalaiah, S. (2025). Sustainability in Cloud: How Enterprises Are Reducing Carbon Footprints. Tech Mahindra. Available at https://www.techmahindra.com/insights/views/sustainability-in-cloud/

[22] Gupta, K., & Katiyar, V. (2016). Energy aware virtual machine migration techniques for cloud environment. International Journal of Computer Applications, 141(2), 11-16.

[23] Han, K., Cai, X., Zhang, X., & Wang, C. (2015, April). Research on energy efficiency evaluation in the cloud. In 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering (pp. 441-445). Atlantis Press.

[24] Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. MIT Technology Review. Available at https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

[25] Hivenet (2025). Who has the greenest cloud? The most sustainable cloud tech in 2026. Available at https://www.hivenet.com/post/who-has-the-greenest-cloud-the-most-sustainable-cloud-tech

[26] Jiang, Y., Roy, R. B., Li, B., & Tiwari, D. (2024, November). Ecolife: Carbon-aware serverless function scheduling for sustainable computing. In SC24: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15). IEEE.

[27] Kabir, M. H., Islam, M. N., Khan, M. R. A. A., Newaz, S. S., & Mahamud, M. S. (2022). Optimizing data center operations with artificial intelligence and machine learning. American Journal of Scholarly Research and Innovation, 1(01), 53-75.

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

[29] Kaur, A., Kaur, B., Singh, P., Devgan, M. S., & Toor, H. K. (2020). Load balancing optimization based on deep learning approach in cloud environment. International Journal of Information Technology and Computer Science, 12(3), 8-18.

[30] Khan, N., Haugerud, H., Shrestha, R., & Yazidi, A. (2019, November). Optimizing power and energy efficiency in cloud computing. In Proceedings of the 11th International Conference on Management of Digital EcoSystems (pp. 256-261).

[31] KHAPRA, C. (2024). A survey on end-to-end speech recognition systems. INTERNATIONAL JOURNAL OF COMMUNICATION, 5(2), 122-127.

[32] Khapra, P. (2022). Evolution of cybercrime and deepfakes–Exploring intervention strategies of international organizations against AI threats. NeuroQuantology, 20(22), 1425-1434.

[33] Kräuter, T. (2024). This Month's Reason Technology will Save the World: Energy Savings and Serverless Principles. Intive. Available at https://www.intive.com/insights/this-months-reason-technology-will-save-the-world-energy-savings-and-serverless-principles

[34] Lavi, H. (2025). Measuring greenhouse gas emissions in data centers: the environmental impact of cloud computing. Climatiq. Available at https://www.climatiq.io/blog/measure-greenhouse-gas-emissions-carbon-data-centres-cloud-computing

[35] Li, J., Guo, Y., Zhang, X., & Fu, Z. (2021). Using hybrid machine learning methods to predict and improve the energy consumption efficiency in oil and gas fields. Mobile Information Systems, 2021(1), 5729630.

[36] Mavridis, I., & Karatza, H. (2019). Combining containers and virtual machines to enhance isolation and extend functionality on cloud computing. Future Generation Computer Systems, 94, 674-696.

[37] Mireslami, S., Rakai, L., Wang, M., & Far, B. H. (2019). Dynamic cloud resource allocation considering demand uncertainty. IEEE Transactions on Cloud Computing, 9(3), 981-994.

[38] Mohamed, A. A., Seyam, E. A., Elsaeed, A. R., Abualigah, L., Smerat, A., AbdelMouty, A. M., & Refaat, H. E. (2025). Energy Aware Task Scheduling of IoT Application Using a Hybrid Metaheuristic Algorithm in Cloud Computing. Computers, Materials & Continua, 86(3).

[39] Mughees, A., Tahir, M., Sheikh, M. A., & Ahad, A. (2020). Towards energy efficient 5G networks using machine learning: Taxonomy, research challenges, and future research directions. Ieee Access, 8, 187498-187522.

[40] Oberhaus, D. (2019). Amazon, Google, Microsoft: Here's Who Has the Greenest Cloud. Wired. Available at https://www.wired.com/story/amazon-google-microsoft-green-clouds-and-hyperscale-data-centers/

[41] Oloruntoba, O. (2024). Green cloud computing: AI for sustainable database management. World Journal of Advanced Research and Reviews, 23(03), 3242-3257.

[42] Pecan (2024). Optimize Efficiency With AI-Driven Energy Management. Available at https://www.pecan.ai/blog/optimize-efficiency-with-ai-energy-management/

[43] Puso, N., Sigwele, T., & Mustapha, O. Z. (2023). Machine learning centered energy optimization in cloud computing: A review. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 11(3), 834-853.

[44] Radhamani, V., & Dalin, G. (2019). Significance of artificial intelligence and machine learning techniques in smart cloud computing: a review. International Journal of Soft Computing and Engineering, 9(3), 1-7.

[45] Rathor, V. S., Pateriya, R. K., & Gupta, R. K. (2015). An efficient virtual machine scheduling technique in cloud computing environment. International Journal of Modern Education and Computer Science, 7(3), 39.

[46] Shrimali, B., & Patel, H. (2015, August). Performance based energy efficient techniques for VM allocation in cloud environment. In proceedings of the third international symposium on women in computing and informatics (pp. 477-486).

[47] Shukla, S., Mishra, S., Singh, J., Mishra, K., Khapra, C., & Morey, P. (2024, August). Deep Belief Networks for Unsupervised Feature Learning and Improved Image Recognition. In International Conference on Artificial Intelligence on Textile and Apparel (pp. 133-145). Singapore: Springer Nature Singapore.

[48] SRILATHA, D. (2019). Machine vs non-machine learning approaches to cloud security solutions: a survey. International journal of engineering and technology.

[49] Srivastava, P., & Meenu, M. (2025). Green Computing: Energy-Efficient AI for Sustainable Cloud Infrastructure. Meenu and Mrs., Meenu, Green Computing: Energy-Efficient AI for Sustainable Cloud Infrastructure (November 12, 2025).

[50] Svystun, O. (2024). How to Achieve Energy Efficiency and Sustainability in Cloud-Based Solutions. Techstack. Available at https://tech-stack.com/blog/how-to-achieve-energy-efficiency-and-sustainability-in-cloud-based-solutions/

[51] Usman, M. J., Ismail, A. S., Abdul-Salaam, G., Chizari, H., Kaiwartya, O., Gital, A. Y., ... & Dishing, S. I. (2019). Energy-efficient nature-inspired techniques in cloud computing datacenters. Telecommunication Systems, 71(2), 275-302.

[52] Vakilinia, S., Heidarpour, B., & Cheriet, M. (2016). Energy efficient resource allocation in cloud computing environments. IEEE Access, 4, 8544-8557.

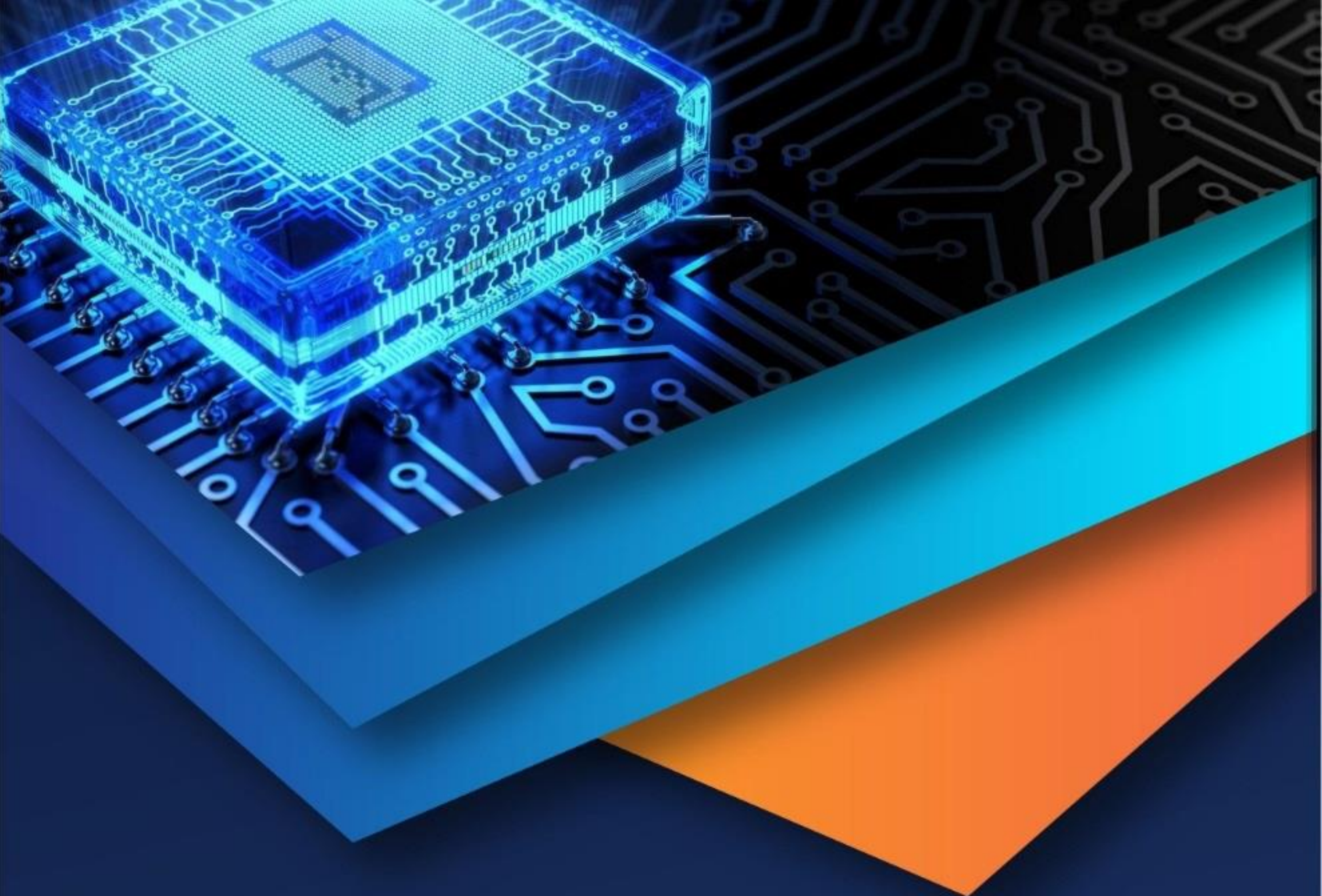
[53] Wengel, G. (2025). A guide to carbon-aware computing. Climatiq. Available at https://www.climatiq.io/blog/a-guide-to-carbon-aware-computing

[54] Wilson, M. (2024). Understanding Power Usage Effectiveness in Sustainable Data Centers. Nlyte Software. Available at https://www.nlyte.com/blog/understanding-power-usage-effectiveness-in-sustainable-data-centers/.

[55] Xaba, X. C., Ogudo, K., & Chabalala, S. (2025, November). Optimizing cloud computing and network resource allocation through cloud automation. In Proceedings of the 2025 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (pp. 1-6).

[56] Yunus, A. (2025). How to Reduce AI Energy Costs with Model Optimization. Artech Digital. Available at https://www.artech-digital.com/blog/how-to-reduce-ai-energy-costs-with-model-optimization

[57] Zhang, J., Chen, B., Zhao, Y., Cheng, X., & Hu, F. (2018). Data security and privacy-preserving in edge computing paradigm: Survey and open issues. IEEE access, 6, 18209-18237.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY