

A REVIEW OF CARBON-AWARE INTELLIGENT SCHEDULING OF MACHINE LEARNING WORKLOADS IN GEOGRAPHICALLY DISTRIBUTED DATA CENTERS

Sachin Kumar Gupta¹, Mrs. Arifa Khan²

¹Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

²Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

Abstract - The rapid expansion of machine learning (ML) applications has significantly increased the computational demand of modern data centers, intensifying their energy consumption and associated carbon emissions. As geographically distributed data centers become the backbone of large-scale AI services, carbon-aware intelligent scheduling has emerged as a promising strategy to mitigate environmental impact while maintaining performance guarantees. This review systematically examines existing approaches to carbon-aware scheduling of ML workloads across geographically distributed infrastructures. We analyze methods that leverage real-time and forecasted grid carbon intensity, renewable energy availability, geographic load balancing, and intelligent optimization techniques such as machine learning-driven prediction and reinforcement learning-based decision making. The review categorizes prior studies based on scheduling objectives, optimization strategies, and evaluation metrics, highlighting trade-offs among carbon reduction, latency, cost, and quality of service. Furthermore, we compare architectural frameworks, workload characteristics, and deployment assumptions to identify common design patterns and limitations. Key challenges—including data uncertainty, scalability constraints, multi-objective conflicts, and lack of standardized benchmarks—are critically discussed. Finally, emerging research directions are outlined, emphasizing integrated AI-driven orchestration, cross-layer optimization, and sustainability-aware system design. By synthesizing current knowledge and identifying research gaps, this review aims to provide a comprehensive reference for researchers and practitioners seeking to design environmentally sustainable ML scheduling strategies in distributed cloud environments.

Key Words: Carbon-aware computing; Machine learning workloads; Geographically distributed data centers; Intelligent scheduling; Renewable energy integration; Sustainable cloud computing.

1. INTRODUCTION

The rapid digital transformation of modern society has led to an unprecedented expansion of cloud services, artificial intelligence (AI), and large-scale data processing systems. While these advancements drive innovation and economic growth, they also contribute to increasing energy demand

and carbon emissions. Data centers, which serve as the backbone of digital infrastructure, are now recognized as significant energy consumers worldwide. Within this context, carbon-aware intelligent scheduling of machine learning (ML) workloads has emerged as a promising strategy to balance computational performance with environmental sustainability. This section introduces the background, motivation, and scope of this review.

1.1 Background

1.1.1 Climate Change and Carbon Emissions from Digital Infrastructure

Global climate change is strongly linked to greenhouse gas emissions, particularly carbon dioxide (CO₂), generated from fossil fuel-based energy systems. The information and communication technology (ICT) sector contributes a non-negligible share of global emissions, driven by increasing internet usage, cloud computing, and AI applications. Recent assessments indicate that digital infrastructure, including networks and data centers, accounts for a growing percentage of global electricity consumption (IEA, 2023). Although improvements in hardware efficiency have partially mitigated growth in energy demand, the overall carbon footprint of computing continues to rise due to expanding service requirements.

Digital platforms increasingly rely on high-performance computing clusters to support data-intensive tasks such as deep learning model training and large-scale analytics. Studies have shown that training large neural networks can produce substantial carbon emissions, especially when powered by carbon-intensive grids (Strubell, Ganesh and McCallum, 2019). Consequently, sustainability has become a critical design objective in next-generation computing systems.

1.1.2 Role of Data Centers in Global Energy Consumption

Data centers form the core infrastructure supporting cloud services, enterprise applications, and AI systems. Their energy consumption arises from computing equipment, storage systems, networking devices, and cooling infrastructure. According to global energy assessments, data

centers consume hundreds of terawatt-hours (TWh) of electricity annually, with demand projected to grow alongside digitalization trends (IEA, 2023).

To evaluate efficiency, metrics such as Power Usage Effectiveness (PUE) have been widely adopted. However, while PUE captures facility-level efficiency, it does not directly reflect carbon intensity, which depends on the regional energy mix (Barroso, Clidas and Hölzle, 2018). Therefore, two data centers with similar efficiency levels may exhibit significantly different carbon footprints depending on their geographic location and access to renewable energy sources.

The geographic distribution of hyperscale data centers introduces both opportunities and complexities. Operators can potentially shift workloads across regions to exploit lower-carbon electricity, yet such strategies must carefully consider latency constraints, network costs, and service-level agreements.

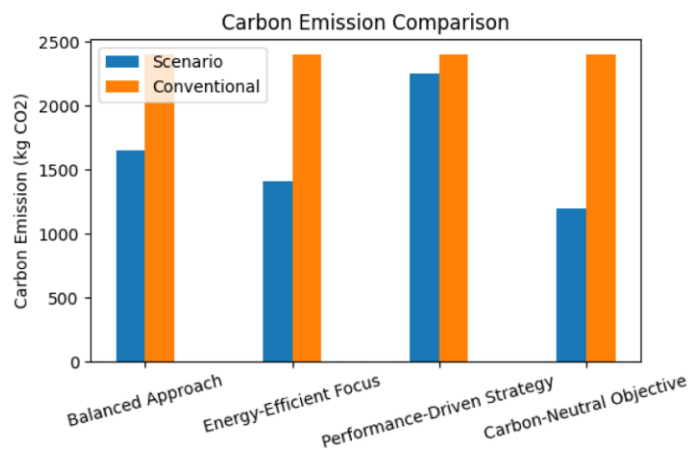


Figure-1: Temporal Carbon Intensity of Electricity Grids

1.1.3 Growth of Machine Learning Workloads and Sustainability Challenges

Machine learning workloads have grown rapidly in both scale and complexity. Modern deep learning models require extensive computational resources for training, often involving distributed GPU or TPU clusters. Research highlights that the computational cost of state-of-the-art AI models has increased exponentially over recent years, leading to corresponding increases in energy consumption (Amodei et al., 2018).

Unlike traditional enterprise workloads, ML tasks—particularly batch training jobs—often exhibit temporal flexibility, making them suitable candidates for carbon-aware scheduling. However, real-time inference services may impose strict latency requirements, limiting relocation or delay strategies. Furthermore, workload heterogeneity, unpredictable demand, and dynamic grid carbon intensity introduce additional scheduling challenges.

These sustainability concerns have motivated research into intelligent orchestration mechanisms capable of aligning ML workload execution with periods or locations of lower carbon intensity, thereby reducing environmental impact without sacrificing performance.

1.2 Scope and Objectives of the Review

1.2.1 Focus on Scheduling Techniques

This review concentrates specifically on carbon-aware scheduling mechanisms rather than hardware design or cooling innovations. Scheduling refers to the allocation of computational workloads across time and geographic locations to optimize defined objectives. In carbon-aware contexts, scheduling decisions integrate real-time or forecasted carbon intensity data, renewable energy availability, and workload characteristics.

Prior work has explored heuristic-based algorithms, optimization models, and machine learning-driven controllers for reducing emissions while preserving quality of service (Qureshi et al., 2009; Radovanović, Konigstein and Schneider, 2022). By synthesizing these approaches, this review aims to classify methodologies, evaluate their strengths and limitations, and identify research gaps.

1.2.2 Importance of Geographic Distribution

Geographically distributed data centers provide a unique opportunity for carbon-aware workload shifting. Since electricity grids differ in carbon intensity depending on generation mix—such as coal, natural gas, hydro, wind, or solar—relocating workloads can significantly reduce associated emissions. Geographic load balancing has been proposed as an effective mechanism to exploit spatial diversity in renewable energy availability (Liu et al., 2011).

However, geographic scheduling must address constraints such as network latency, bandwidth limitations, data sovereignty regulations, and migration overhead. The interplay between carbon efficiency and service performance represents a central theme in the literature and forms a core focus of this review.

1.2.3 Relevance to Sustainable Computing

Sustainable computing extends beyond energy efficiency to encompass broader environmental and societal impacts. Carbon-aware intelligent scheduling aligns with global sustainability goals by enabling data center operators to actively reduce emissions through informed operational decisions. As regulatory frameworks and corporate sustainability commitments increasingly emphasize carbon accounting and reporting, integrating carbon awareness into workload management becomes strategically important.

By systematically reviewing carbon-aware intelligent scheduling strategies for ML workloads in distributed

environments, this paper aims to provide a consolidated foundation for advancing environmentally responsible cloud and AI systems.

2. FUNDAMENTALS AND KEY CONCEPTS

Understanding carbon-aware intelligent scheduling requires a clear foundation in machine learning workload characteristics, carbon-aware computing principles, distributed data center architectures, and the metrics used to evaluate sustainability and performance. This section outlines the essential technical and conceptual elements that underpin the research domain.

2.1 Machine Learning Workloads in Data Centers

Machine learning workloads have become a dominant component of cloud computing environments, particularly with the widespread adoption of deep learning models across domains such as computer vision, natural language processing, and recommendation systems. These workloads exhibit unique computational patterns and resource demands compared to traditional enterprise applications.

2.1.1 Workload Types: Training vs Inference

ML workloads in data centers are broadly categorized into training and inference tasks. Training involves iterative optimization over large datasets to update model parameters and typically requires substantial computational power over extended durations. Distributed training frameworks often rely on clusters of GPUs or TPUs interconnected with high-speed networks to reduce training time (Li et al., 2020). As model sizes increase, the energy and carbon footprint of training large neural networks has become a major sustainability concern (Strubell, Ganesh and McCallum, 2019).

Inference workloads, in contrast, deploy trained models to generate predictions in real time or near real time. These tasks are generally latency-sensitive and must meet strict service-level objectives. Unlike training, inference workloads are more continuous and user-driven, often requiring geographically closer deployment to end users to minimize response times. The distinction between training and inference is critical for carbon-aware scheduling because training jobs are typically more flexible in time and location, whereas inference tasks are more constrained.

2.1.2 Resource Requirements and Variability

ML workloads demand heterogeneous resources, including CPUs, GPUs, high-bandwidth memory, storage systems, and high-speed interconnects. Deep learning training is particularly resource-intensive due to large model sizes and massive datasets. Resource consumption can vary significantly depending on model architecture, batch size, and dataset characteristics (Amodei et al., 2018).

Moreover, workload demand in data centers fluctuates over time due to varying user activity patterns and business requirements. This variability complicates scheduling decisions, as operators must dynamically allocate resources while balancing performance, cost, and environmental impact. Effective carbon-aware scheduling mechanisms must therefore account for workload elasticity, heterogeneity, and temporal flexibility.

2.2 Carbon Awareness in Computing

Carbon-aware computing integrates environmental considerations into system design and operational decision-making. Rather than focusing solely on energy efficiency, it emphasizes reducing carbon emissions associated with electricity consumption.

2.2.1 Concept of Carbon-Aware Computing

Carbon-aware computing refers to strategies that adjust computational activities based on the carbon intensity of the electricity grid. Carbon intensity varies spatially and temporally depending on the mix of generation sources, such as coal, natural gas, nuclear, hydro, wind, or solar. By aligning workload execution with periods or regions of lower carbon intensity, systems can reduce overall emissions without necessarily reducing total energy consumption (Radovanović, Konigstein and Schneider, 2022).

This paradigm often relies on real-time or forecasted carbon intensity data provided by grid operators or external monitoring services. Scheduling systems use this information to determine when and where workloads should run to minimize carbon impact.

2.2.2 Importance for Data Center Operations

For data center operators, carbon awareness is increasingly tied to regulatory compliance, corporate sustainability commitments, and environmental reporting frameworks. While traditional efficiency metrics such as Power Usage Effectiveness (PUE) improve infrastructure efficiency, they do not capture the carbon intensity of electricity sources (Barroso, Clidaras and Hölzle, 2018). Therefore, two equally efficient facilities may have vastly different environmental impacts depending on their geographic location.

Incorporating carbon awareness into scheduling decisions enables operators to reduce Scope 2 emissions associated with purchased electricity. This shift represents a move from static efficiency optimization to dynamic, environmentally informed orchestration.

2.3 Geographically Distributed Data Centers

Geographically distributed data centers consist of multiple facilities located across different regions or countries, interconnected through high-speed networks. These architectures are common among hyperscale cloud providers to ensure redundancy, scalability, and low-latency service delivery.

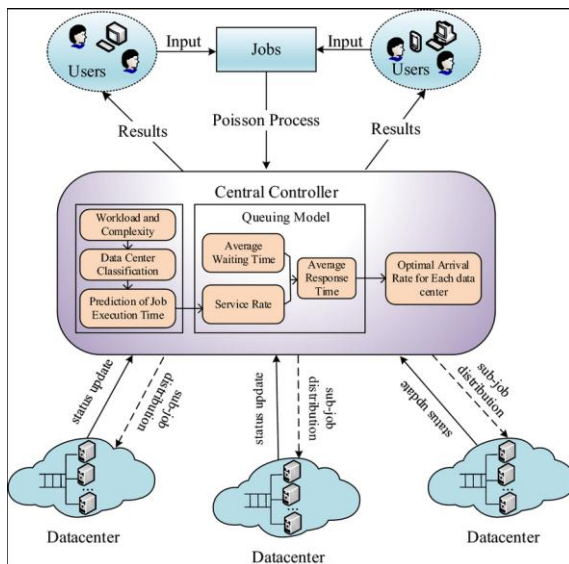


Figure-2: Geographic Load Balancing Concept

2.3.1 Definitions and Architectures

Distributed data center architectures typically follow a hierarchical or federated model, where workloads can be allocated across regions based on resource availability and demand. Such infrastructures leverage virtualization and containerization technologies to enable flexible workload migration and replication (Beloglazov and Buyya, 2012).

The geographic dispersion of facilities exposes them to diverse electricity markets and energy mixes. As a result, carbon intensity can vary substantially between locations at any given time, creating opportunities for spatial load shifting.

2.3.2 Benefits and Challenges in Workload Scheduling

The primary benefit of geographically distributed systems is the ability to perform geographic load balancing—redirecting computational tasks to locations with lower operational cost or carbon intensity. Research has demonstrated that such strategies can reduce energy costs and emissions while maintaining service quality under certain constraints (Qureshi et al., 2009).

However, workload migration introduces challenges. Data transfer latency, bandwidth limitations, regulatory constraints on data locality, and potential service disruptions must be carefully managed. For latency-sensitive inference

services, excessive relocation may degrade user experience. Therefore, intelligent scheduling mechanisms must evaluate trade-offs among carbon reduction, cost, and performance.

2.4 Metrics and Evaluation Criteria

Evaluating carbon-aware scheduling strategies requires multidimensional performance metrics that capture both environmental and operational objectives.

2.4.1 Carbon Emissions (CO₂eq)

Carbon emissions are typically measured in kilograms or metric tons of carbon dioxide equivalent (CO₂eq), accounting for various greenhouse gases standardized to CO₂ impact. In data centers, emissions are often calculated by multiplying electricity consumption by the carbon intensity factor of the regional grid (IEA, 2023). Accurate measurement depends on reliable carbon intensity data and transparent accounting methodologies.

2.4.2 Power Usage Effectiveness (PUE)

Power Usage Effectiveness (PUE) is defined as the ratio of total facility energy consumption to IT equipment energy consumption. It provides an indicator of infrastructure efficiency, particularly cooling and power distribution systems. While widely adopted, PUE does not directly reflect environmental impact unless combined with carbon intensity metrics (Barroso, Clidaras and Hölzle, 2018). Consequently, carbon-aware scheduling research increasingly supplements PUE with emission-based indicators.

2.4.3 Latency, Throughput, and Energy Efficiency

Operational performance metrics remain essential. Latency measures response time for service requests, which is critical for real-time ML inference. Throughput reflects the volume of tasks processed within a given period, while energy efficiency assesses computational output per unit of energy consumed. Carbon-aware scheduling solutions must balance these performance indicators against emission reduction objectives to ensure practical deployability.

3. REVIEW OF CARBON-AWARE SCHEDULING TECHNIQUES

Carbon-aware scheduling has evolved from simple energy-cost minimization strategies to sophisticated, intelligence-driven frameworks that incorporate carbon intensity signals, renewable energy availability, and workload flexibility. This section critically reviews the major thematic categories of existing approaches, highlighting their methodological evolution, strengths, and limitations.

3.1 Scheduling Based on Real-Time Carbon Intensity

Early carbon-aware scheduling research focused on leveraging spatial and temporal variations in grid carbon intensity to reduce emissions.

3.1.1 Techniques Leveraging Grid Carbon Intensity Forecasts

Several studies propose adjusting workload execution based on real-time or forecasted carbon intensity data from electricity grids. By deferring or relocating flexible tasks to periods of lower carbon intensity, data centers can reduce indirect emissions without modifying hardware infrastructure. For example, carbon-intensity-aware workload shifting mechanisms have been implemented in large-scale production environments, demonstrating measurable emission reductions while maintaining service-level objectives (Radovanović, Koningstein and Schneider, 2022).

Forecast-driven approaches typically integrate time-series prediction models to anticipate short-term carbon fluctuations. While effective in reducing emissions, their performance depends heavily on forecast accuracy and workload flexibility. Sudden changes in grid conditions may limit optimization potential.

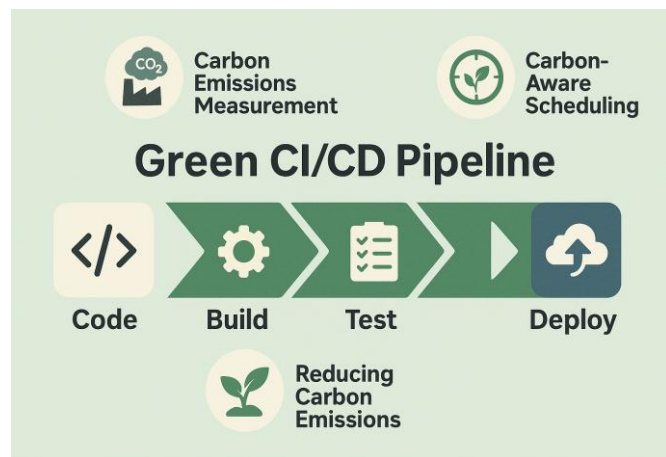


Figure-3: Carbon-Aware Scheduling Workflow

3.1.2 Carbon Pricing and Power Source Mix-Based Scheduling

Other approaches incorporate carbon pricing or electricity market signals into scheduling decisions. By internalizing carbon cost into operational objectives, these models align environmental goals with economic incentives. Research on geographically distributed systems has shown that incorporating regional electricity prices and energy mix information can significantly reduce both operational cost and emissions (Liu et al., 2011).

However, such strategies may prioritize cost savings over strict carbon minimization, especially when low-carbon electricity is expensive. Moreover, reliance on market mechanisms introduces regulatory and regional variability, limiting universal applicability.

Advantages and Limitations:

Real-time carbon-aware scheduling is relatively straightforward to integrate into existing cloud orchestration systems and does not require complex predictive modeling. Nonetheless, its effectiveness is constrained by workload inflexibility, limited migration capacity, and uncertainty in carbon data streams.

3.2 Predictive and Intelligent Scheduling

As workloads and grid dynamics became more complex, research shifted toward predictive and AI-driven scheduling mechanisms.

3.2.1 ML-Based Forecasting for Emissions and Workload Demand

Predictive models are increasingly used to estimate both future carbon intensity and workload demand. Machine learning techniques such as regression models and deep neural networks can capture nonlinear temporal patterns in grid emissions and user activity. Surveys of distributed machine learning systems highlight the importance of predictive resource allocation to improve efficiency and scalability (Li et al., 2020).

By combining carbon forecasts with workload demand predictions, schedulers can proactively allocate tasks to optimal locations and times. However, model training overhead and forecasting errors remain critical concerns.

3.2.2 Reinforcement Learning and Heuristic Algorithms

Reinforcement learning (RL) has gained attention for adaptive decision-making in dynamic cloud environments. RL-based schedulers learn optimal policies through interaction with system states, balancing carbon reduction and performance objectives. Compared to static heuristics, RL approaches can adapt to changing grid conditions and workload patterns.

Heuristic algorithms, including greedy and metaheuristic methods, have also been widely adopted due to their computational efficiency. Energy-aware dynamic consolidation techniques in virtualized environments illustrate how heuristic approaches can achieve practical emission savings with manageable complexity (Beloglazov and Buyya, 2012).

3.2.3 Adaptive Scheduling Under Uncertainty

Uncertainty in carbon intensity forecasts, renewable generation, and workload arrivals necessitates robust scheduling frameworks. Adaptive methods incorporate stochastic optimization or online learning to respond to unexpected fluctuations. Although adaptive systems improve resilience, they introduce additional computational overhead and design complexity.

3.3 Geographic Load Balancing and Migration

Geographic load balancing represents one of the most influential paradigms in carbon-aware computing.

3.3.1 Distributing Workloads Across Multiple Sites

By distributing workloads across geographically dispersed data centers, operators can exploit regional differences in carbon intensity. Foundational research demonstrated that shifting workloads to locations with lower electricity costs or cleaner energy mixes can substantially reduce overall energy expenditure and emissions (Qureshi et al., 2009).

Such approaches are particularly effective for delay-tolerant batch workloads, including ML model training tasks. However, large-scale migration requires robust network connectivity and data replication mechanisms.

3.3.2 Policies for Shifting Tasks to Low-Carbon Regions

Policy-based frameworks determine when and where workloads should be relocated. Policies may consider thresholds in carbon intensity, forecast confidence intervals, or hybrid cost-carbon objectives. Empirical studies have shown that strategic geographic shifting can reduce emissions without significant performance degradation, provided that latency-sensitive tasks are carefully managed (Liu et al., 2011).

3.3.3 Trade-offs: Energy vs Latency

A central challenge in geographic scheduling lies in balancing energy or carbon reduction against latency constraints. Real-time inference services often require proximity to end users, limiting relocation options. Excessive migration may also increase network energy consumption, partially offsetting carbon savings. Therefore, effective strategies must carefully evaluate system-wide impacts rather than focusing solely on computational emissions.

3.4 Renewable Energy Integration

Integration of renewable energy sources has become a key dimension of carbon-aware scheduling research.

3.4.1 Aligning Workload Timing with Renewable Availability

Temporal alignment strategies schedule flexible workloads during periods of high renewable generation, such as midday solar peaks or strong wind conditions. Research on sustainable data center design emphasizes the importance of workload deferral and energy-aware orchestration to maximize renewable utilization (Barroso, Clidaras and Hölzle, 2018).

This approach is particularly suitable for ML training tasks that can tolerate execution delays. However, renewable intermittency limits reliability without complementary mechanisms.

3.4.2 Hybrid Scheduling with Energy Storage Systems

Energy storage systems, such as battery arrays, enhance renewable integration by smoothing variability. Hybrid scheduling models coordinate workload allocation with both renewable generation forecasts and storage capacity constraints. While such systems can significantly reduce carbon intensity, they introduce additional capital and operational costs, complicating economic feasibility assessments.

3.5 Multi-Objective and Optimization Frameworks

Carbon-aware scheduling is inherently a multi-objective problem involving trade-offs among emissions, cost, and performance.

3.5.1 Optimization Methods Considering Carbon and Performance

Mathematical optimization techniques—including linear programming, mixed-integer programming, and convex optimization—have been used to formalize carbon-performance trade-offs. These models often incorporate constraints such as service-level agreements, resource capacities, and migration overhead.

Optimization-based frameworks provide theoretical guarantees but may face scalability challenges in large hyperscale environments.

3.5.2 Pareto-Based and Constraint-Driven Solutions

Pareto optimization approaches identify a set of non-dominated solutions balancing competing objectives. Such methods allow operators to select preferred trade-off points depending on strategic priorities. Constraint-driven formulations further ensure compliance with strict latency or reliability requirements while minimizing carbon emissions. Although computationally intensive, these frameworks offer comprehensive decision-support capabilities.

3.6 Emerging Trends and AI-Assisted Frameworks

Recent research extends carbon-aware scheduling beyond centralized cloud architectures.

3.6.1 Edge Computing Integration

The proliferation of edge computing introduces new opportunities and complexities. Deploying ML inference closer to end users reduces latency but may increase total energy consumption due to smaller-scale infrastructure. Integrating carbon-aware mechanisms across cloud-edge hierarchies requires coordinated orchestration strategies and distributed decision-making frameworks.

3.6.2 Federated and Decentralized Scheduling Approaches

Federated and decentralized models distribute control across multiple administrative domains, enhancing scalability and resilience. Such architectures reduce single-point decision bottlenecks and can integrate localized carbon signals. However, coordination overhead and inconsistent policy enforcement remain open challenges.

4. COMPARATIVE ANALYSIS

A comparative analysis is essential in a review paper to move beyond descriptive summarization and provide structured evaluation of prior studies. This section synthesizes the reviewed literature by comparing methodologies, evaluation metrics, carbon reduction effectiveness, performance trade-offs, and real-world applicability. Rather than treating each study in isolation, the discussion identifies patterns, strengths, and limitations across thematic categories.

4.1 Tabular Comparison of Reviewed Studies

A tabular synthesis of the literature enables systematic comparison across multiple dimensions, including scheduling methodology, optimization objective, evaluation metrics, reported carbon savings, performance impact, and scalability. Comparative frameworks are commonly used in cloud computing surveys to identify methodological trends and practical limitations (Beloglazov and Buyya, 2012).

4.1.1 Methodology

The reviewed studies employ diverse methodological approaches, including heuristic-based scheduling, mathematical optimization, geographic load balancing, predictive modeling, and reinforcement learning. Early work largely relied on deterministic or cost-driven optimization models focusing on electricity price and energy consumption (Qureshi et al., 2009). More recent research integrates carbon intensity signals and AI-based decision mechanisms to dynamically adapt to grid and workload variability (Radovanović, Koningstein and Schneider, 2022).

Heuristic approaches are typically lightweight and computationally efficient but may lack global optimality guarantees. Optimization-based models provide theoretical rigor but often struggle with scalability in hyperscale cloud environments. Learning-based methods demonstrate adaptability but require training data, computational overhead, and careful tuning.

4.1.2 Metrics Used

Across studies, evaluation metrics vary significantly. Carbon emissions are commonly measured in CO₂ equivalent (CO₂eq), calculated using regional carbon intensity factors. Some research emphasizes energy consumption alone, while others integrate multi-dimensional metrics such as latency, throughput, operational cost, and Power Usage Effectiveness (PUE) (Barroso, Clidaras and Hölzle, 2018).

The absence of standardized benchmarks complicates cross-study comparison. For example, some works evaluate carbon reduction percentage relative to baseline scheduling, whereas others report absolute emission values. Additionally, workload assumptions differ, ranging from synthetic traces to real-world production workloads, limiting generalizability.

4.1.3 Carbon Savings Achieved

Reported carbon savings vary widely depending on workload flexibility and geographic diversity. Geographic load balancing studies demonstrate notable emission reductions when workloads can be shifted across regions with heterogeneous energy mixes (Liu et al., 2011). Similarly, carbon-intensity-aware scheduling in production-scale systems has shown measurable reductions without major performance degradation (Radovanović, Koningstein and Schneider, 2022).

However, savings are often context-dependent. Delay-tolerant ML training workloads yield higher reduction potential compared to latency-sensitive inference services. Furthermore, carbon reduction gains diminish in regions where the grid is already dominated by low-carbon energy sources.

4.1.4 Performance Impacts

Performance trade-offs remain central to comparative evaluation. While some approaches achieve carbon reductions with minimal latency impact, aggressive workload migration can increase response times or network overhead. Studies on distributed systems emphasize that cost and carbon optimization must not compromise service-level agreements (Qureshi et al., 2009).

Learning-based scheduling can improve adaptability but may introduce decision latency or computational overhead. In large-scale ML systems, performance degradation during

migration or task deferral must be carefully balanced against environmental benefits.

4.1.5 Scalability and Practical Applicability

Scalability is a recurring concern. Heuristic and rule-based approaches generally scale well and are easier to integrate into existing cloud orchestration frameworks. Optimization-based models may become computationally expensive as the number of data centers and workloads increases. Production-oriented carbon-aware scheduling frameworks highlight the importance of incremental deployment strategies and compatibility with existing cloud management systems (Radovanović, Koningstein and Schneider, 2022).

Practical applicability also depends on data availability, including real-time carbon intensity signals and accurate workload forecasting. Many academic studies rely on simulated environments, which may not fully capture real-world operational complexity.

4.2 Synthesizing Insights

Beyond tabular comparison, synthesizing insights reveals broader trends and research gaps within the field.

4.2.1 What Works Best Under Different Conditions?

The effectiveness of carbon-aware scheduling strategies depends largely on workload characteristics and infrastructure heterogeneity. Geographic load balancing performs best when significant spatial variation in carbon intensity exists and workloads are delay-tolerant. Predictive and reinforcement learning approaches are more suitable in highly dynamic environments where both workload demand and grid carbon intensity fluctuate frequently.

Renewable-aligned scheduling is particularly effective in regions with high solar or wind penetration. However, in grids dominated by stable low-carbon sources such as hydro or nuclear, marginal benefits may be limited. Therefore, context-aware hybrid strategies often outperform single-objective methods.

4.2.2 Gaps in Current Research

Several research gaps remain evident. First, there is limited availability of standardized benchmarking frameworks for carbon-aware ML scheduling. This restricts reproducibility and fair comparison across studies. Second, most research assumes homogeneous administrative control over geographically distributed data centers, whereas real-world cloud ecosystems often involve multi-tenant and multi-provider constraints.

Additionally, carbon accounting methodologies vary, and few studies incorporate full life-cycle assessment of infrastructure. Integration of carbon-aware scheduling with

emerging paradigms such as edge computing and federated learning also remains underexplored.

4.2.3 Conflicting Findings and Debates

Some debates persist regarding the trade-off between carbon reduction and energy efficiency. While energy-efficient systems often reduce emissions, this is not universally true in regions with low-carbon electricity. Moreover, shifting workloads geographically may increase network energy consumption, partially offsetting gains from cleaner grids.

Another area of debate concerns the relative superiority of optimization-based versus learning-based approaches. Optimization models offer interpretability and theoretical guarantees, whereas AI-driven methods provide adaptability but may lack transparency. The literature suggests that hybrid frameworks combining predictive modeling with constrained optimization may offer a balanced solution.

5. CHALLENGES AND OPEN ISSUES

Despite significant progress in carbon-aware intelligent scheduling for machine learning workloads, several technical, operational, and regulatory challenges remain unresolved. These challenges limit large-scale adoption and highlight important directions for future investigation.

5.1 Data Availability and Accuracy of Carbon Intensity Modeling

Carbon-aware scheduling fundamentally depends on accurate, high-resolution carbon intensity data. However, the availability and granularity of such data vary significantly across regions. In many electricity markets, real-time carbon intensity information is either unavailable or published with limited temporal resolution. Furthermore, discrepancies may arise between average grid intensity values and marginal emission factors, which more accurately reflect the emissions impact of incremental electricity demand (Hawkes, 2010).

Forecasting carbon intensity introduces additional uncertainty. Prediction errors—caused by unexpected renewable variability, grid congestion, or sudden demand spikes—can reduce the effectiveness of scheduling decisions. While machine learning models can improve forecasting accuracy, their performance depends on high-quality historical datasets and stable grid behavior. Inaccurate modeling may lead to suboptimal decisions, undermining carbon reduction goals.

5.2 Trade-offs Between Sustainability and Performance

A persistent challenge in carbon-aware scheduling lies in balancing environmental objectives with quality of service requirements. Latency-sensitive ML inference services often

require geographic proximity to end users, limiting the feasibility of workload relocation. Excessive migration or deferral of workloads may increase response times and violate service-level agreements (Qureshi et al., 2009).

Moreover, carbon minimization strategies can conflict with energy efficiency or operational cost objectives. For example, shifting workloads to low-carbon regions may increase network energy consumption or operational expenses. Studies on large-scale AI workloads also emphasize that aggressive optimization may introduce computational overhead or instability in distributed training systems (Strubell, Ganesh and McCallum, 2019). Therefore, effective frameworks must carefully balance multiple objectives rather than optimizing carbon emissions in isolation.

5.3 Scalability and Real-World Deployment Barriers

Many carbon-aware scheduling solutions are validated in simulated or small-scale environments, raising concerns about scalability in hyperscale cloud infrastructures. As the number of data centers, workloads, and constraints increases, optimization problems become computationally complex. Mixed-integer and multi-objective optimization models, while theoretically robust, may not scale efficiently in real-time production settings (Beloglazov and Buyya, 2012).

Practical deployment also requires integration with existing orchestration platforms, container management systems, and workload schedulers. Production-oriented implementations demonstrate feasibility but highlight the importance of incremental adoption and compatibility with operational workflows (Radovanović, Koningstein and Schneider, 2022). Additionally, inter-data center bandwidth limitations and data replication overhead may restrict large-scale workload migration.

5.4 Economic and Policy Considerations

Economic incentives strongly influence adoption of carbon-aware scheduling. Electricity pricing structures, carbon taxation mechanisms, and renewable energy credits vary across jurisdictions. In some markets, low-carbon electricity may be more expensive, discouraging purely carbon-driven workload migration. Incorporating carbon cost into operational decision-making requires transparent pricing mechanisms and standardized reporting frameworks.

Policy and regulatory constraints further complicate geographic scheduling. Data sovereignty laws may restrict cross-border data movement, limiting workload relocation options. Moreover, corporate sustainability reporting standards are still evolving, leading to inconsistencies in carbon accounting methodologies (IEA, 2023). Aligning technical optimization strategies with regulatory compliance remains a significant open issue.

5.5 Security, Privacy, and Fairness Concerns

Carbon-aware workload shifting can introduce new security and privacy challenges. Transferring data across geographic regions may expose systems to additional attack surfaces or compliance risks. Multi-tenant cloud environments must ensure that carbon optimization does not compromise isolation guarantees or data confidentiality.

Fairness considerations also emerge in shared infrastructure. If carbon-aware policies prioritize certain workloads over others, resource allocation inequalities may arise. Furthermore, regions with cleaner energy grids could disproportionately benefit from increased computational activity, raising broader questions about equitable infrastructure utilization. Addressing these concerns requires transparent policy design, robust security frameworks, and fairness-aware scheduling mechanisms.

6. CONCLUSION

This review has systematically examined carbon-aware intelligent scheduling strategies for machine learning workloads in geographically distributed data centers. The analysis highlights a clear evolution from energy-centric optimization models to dynamic, carbon-intensity-aware and AI-driven scheduling frameworks. Real-time carbon-aware workload shifting, predictive modeling, reinforcement learning-based orchestration, renewable-aligned scheduling, and multi-objective optimization approaches collectively demonstrate significant potential to reduce operational emissions while maintaining performance guarantees. Geographic load balancing emerges as a particularly impactful strategy when spatial variation in grid carbon intensity exists and workloads exhibit temporal flexibility.

However, the effectiveness of these approaches depends heavily on accurate carbon intensity data, workload characteristics, infrastructure heterogeneity, and operational constraints. Latency-sensitive inference services, regulatory limitations, and migration overhead often restrict carbon optimization opportunities. The review also reveals a lack of standardized benchmarking frameworks, making cross-study comparisons challenging.

Carbon-aware intelligent scheduling represents a promising pathway toward sustainable AI infrastructure. Future progress will require integrated frameworks that combine predictive analytics, adaptive optimization, and cross-layer coordination across cloud and edge environments. By synthesizing current research trends, identifying methodological strengths, and highlighting persistent challenges, this review provides a consolidated foundation for advancing environmentally responsible scheduling strategies in large-scale distributed computing systems.

6.1. Limitations of the Review

This review is subject to certain limitations. First, the analysis primarily focuses on carbon-aware scheduling at the operational level and does not extensively address hardware-level innovations, cooling technologies, or embodied carbon in infrastructure manufacturing. Second, variations in evaluation methodologies and metrics across studies limit direct quantitative comparison of reported carbon savings. Third, while efforts were made to include representative and influential research, the rapidly evolving nature of this domain means that some recent developments may not be fully captured. Finally, most reviewed studies rely on simulated environments or controlled experiments, which may not fully reflect real-world production complexities.

REFERENCES

1. Amodei, D. et al. (2018) 'AI and compute', OpenAI Blog.
2. Barroso, L.A., Clidaras, J. and Hölzle, U. (2018) *The Datacenter as a Computer: Designing Warehouse-Scale Machines*. 3rd edn. San Rafael: Morgan & Claypool.
3. Beloglazov, A. and Buyya, R. (2012) 'Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers', *Concurrency and Computation: Practice and Experience*, 24(13), pp. 1397–1420.
4. Hawkes, A.D. (2010) 'Estimating marginal CO₂ emissions rates for national electricity systems', *Energy Policy*, 38(10), pp. 5977–5987.
5. IEA (2023) *Data Centres and Data Transmission Networks*. Paris: International Energy Agency.
6. Li, M. et al. (2020) 'A survey of distributed machine learning', *ACM Computing Surveys*, 53(2), pp. 1–36.
7. Liu, Z. et al. (2011) 'Greening geographical load balancing', *Proceedings of the ACM SIGCOMM Conference*.
8. Qureshi, A. et al. (2009) 'Cutting the electric bill for Internet-scale systems', *Proceedings of ACM SIGCOMM*.
9. Radovanović, A., Koningstein, R. and Schneider, I. (2022) 'Carbon-aware computing for datacenters', *IEEE Micro*, 42(4), pp. 50–57.
10. Strubell, E., Ganesh, A. and McCallum, A. (2019) 'Energy and policy considerations for deep learning in NLP', *Proceedings of ACL*, pp. 3645–3650.
11. Alex, M., Ojo, S.O. and Awuor, F.M. (2025) 'Carbon-aware, energy-efficient, and SLA-compliant virtual machine placement in cloud data centers using Deep Q-Networks and agglomerative clustering', *Computers*, 14(7), 280.
12. Bostandoost, R. et al. (2024) 'LACS: learning-augmented algorithms for carbon-aware resource scaling with uncertain demand', arXiv.
13. Guillen-Perez, A. et al. (2025) 'DCcluster-Opt: benchmarking dynamic multi-objective optimization for geo-distributed data center workloads', arXiv.
14. Saad, Z. et al. (2025) 'Towards carbon-aware container orchestration: predicting workload energy consumption with federated learning', arXiv.
15. Breukelman, E. et al. (2024) 'Carbon-aware computing in a network of data centers: a hierarchical game-theoretic approach', arXiv.
16. Danach, K. et al. (2026) 'Carbon-Aware Scheduling in Cloud Computing Operations: a multi-objective optimisation approach', *IET Smart Grid*.
17. Lin, W.T., Chen, X. and Li, Y. (2023) 'Carbon-aware load balance control of data centers with renewable generations', *IEEE/Computer Society*.
18. Zhou, Z. et al. (2013) 'Carbon-Aware Load Balancing for Geo-distributed Cloud Services', *ACM MASCOTS*.
19. Khodayarseresht, E., Shamel, S., Sendi, A., Fournier, Q. and Dagenais, M. (2023) 'Energy and carbon-aware initial VM placement in geographically distributed cloud data centers', *Sustainable Computing*.
20. Zhao, D. et al. (2022) 'An energy and carbon-aware algorithm for renewable-supported virtual machine placement', *Expert Systems with Applications*.
21. Piontek, T., Haghshenas, K. & Aiello, M. (2024) 'Carbon emission-aware job scheduling for Kubernetes deployments', *Journal of Supercomputing*.
22. Xu, K. et al. (2025) 'GREEN: carbon-efficient resource scheduling for machine learning clusters', *Proc. USENIX NSDI* 25.
23. Maji, D. et al. (2023) 'Bringing carbon awareness to multi-cloud application delivery', *Proc. HotCarbon 2023*.
24. Sarkar, S. et al. (2025) 'Hierarchical multi-agent framework for carbon-efficient liquid-cooled data center clusters', arXiv.

25. Sarkar, S. et al. (2024) 'Carbon-aware spatio-temporal workload distribution in cloud data center clusters using reinforcement learning', *ClimateChange.ai / NeurIPS 2024*.
26. Panwar, S.S., Rauthan, M.M.S. & Barthwal, V. (2022) 'A systematic review on effective energy utilization management strategies in cloud data centers', *Journal of Cloud Computing*.
27. Lin, W. et al. (2024) 'A systematic review of green-aware management techniques for sustainable data centers', *Sustainable Computing*.
28. Janani, B. & Deepankumar, E. (2024) 'Improving energy cost efficiency for multiple cloud data centers using green computing', *International Journal of Intelligent Systems and Applications in Engineering*.
29. Gupta, N. & Tyagi, U. (2026) 'Carbon-aware machine learning: designing low-energy AI algorithms for sustainable computing', *IERJ*.
30. Miao, Z. et al. (2024) 'Energy and carbon-aware distributed machine learning tasks scheduling scheme for the multi-renewable energy-based edge-cloud continuum', *Science and Technology for Energy Transition*.
31. Anasuri, S. & Pappula, K.K. (2023) 'Green HPC: carbon-aware scheduling in cloud data centers', *Int. Journal of Emerging Research in Engineering and Technology*.
32. Wiesner, P. et al. (2025) 'Carbon-aware quality adaptation for energy-intensive services', *Proc. ACM E-ENERGY 25*.