Digital Transformation and Technology Dynamics

Journal Homepage: www.techdynamicsjournal.com | ARTICLE NO. DTTD2025002 | VOL. 5 ISSUE 1

Green ICT and Sustainable Computing: Evaluating Carbon Footprint Metrics in Cloud-to-Edge AI Workloads

Rashadul Islam Samrat[1]*, Md Rasel Ul Alam[2], Kanita Haider[3],

[1] *Department of Marketing, University of Barishal, Barishal, Bangladesh*

[2] *Department of Computer and Information Sciences, Williamsburg, Kentucky, USA*

[3] *Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh*

*Corresponding author: samrat.meazi1@gmail.com

Received 19 January 2025; Revised 12 April 2025; Accepted 30 May 2025; Published: 29 July 2025

KEYWORDS

Green ICT,
Sustainable Computing,
Carbon Footprint Metrics, Cloud-to-Edge AI,
Renewable Energy Optimization,
Workload Scheduling,
Grid Carbon Intensity

Abstract

The exponential expansion of Artificial Intelligence (AI) model training and inference workloads across cloud-to-edge infrastructures has led to staggering energy consumption and operational carbon emissions. While cloud datacenters and industrial edge nodes leverage power usage effectiveness (PUE) metrics, traditional workload scheduling algorithms remain carbon-oblivious, optimizing solely for execution cost or completion latency. This paper introduces the Dynamic Carbon-Aware AI Workload Scheduler (DCAS), a mathematical optimization model designed to minimize operational carbon footprints (gCO2eq) across geo-distributed cloud datacenters and industrial edge processing clusters. DCAS formulates a multi-objective Mixed-Integer Linear Programming (MILP) model that dynamically shifts compute-heavy Deep Learning (DL) training and batch inference jobs in space (spatial routing to regions with cleaner energy grids) and time (temporal deferral to match peak solar and wind availability). Evaluating 1,000 deep learning training jobs across 10 geo-distributed cloud and edge facility nodes over a 30-day simulation window using real-time grid carbon intensity feeds, empirical results show that DCAS reduces total operational carbon emissions by 58.4% compared to static cost-optimized baselines, while restricting average Service Level Agreement (SLA) delay penalties to less than 4.2%. Furthermore, spatial-temporal workload shifting paired with edge battery energy storage systems (BESS) reduces peak grid power demand by 42.1%, delivering a scalable blueprint for sustainable Green ICT infrastructure in Industry 4.0 environments.

1. Introduction

The rapid convergence of Large Language Models (LLMs), deep learning, computer vision, Internet of Things (IoT) platforms, and autonomous agentic artificial intelligence (AI) systems is fundamentally transforming modern cloud and edge-computing infrastructures. Contemporary AI applications increasingly depend on computationally intensive model training, fine-tuning, large-scale data processing, and high-frequency inference. These workloads are predominantly executed on GPU- and accelerator-rich data-center infrastructures, where increasing model size, training datasets, and inference demand have resulted in substantial growth in electricity consumption. The International Energy Agency (IEA, 2025) estimates that global data-center electricity consumption reached approximately 415 TWh in 2024, equivalent to around 1.5% of global electricity consumption. Under its base-case scenario, this demand could more than double to approximately 945 TWh by 2030, with AI-accelerated servers representing a major contributor to the projected increase.

The environmental implications of this expansion are becoming increasingly significant. Earlier research has demonstrated that the computational requirements of modern deep learning can translate into substantial energy consumption and associated carbon emissions during model training and experimentation (Strubell et al., 2019). The emergence of increasingly large AI models has intensified this concern, creating a growing tension between improvements in model capability and the environmental cost of computation. Schwartz et al. (2020) therefore introduced the concept of Green AI, emphasizing that computational efficiency and environmental cost should be treated as important evaluation dimensions alongside predictive performance.

At the broader ICT-sector level, estimates indicate that information and communication technologies already account for a significant fraction of global greenhouse-gas (GHG) emissions. A comprehensive assessment by Freitag et al. (2021) estimated the ICT sector's direct and indirect contribution at approximately 1.8%–2.8% of global GHG emissions, with the potential estimate increasing to approximately 2.1%–3.9% when supply-chain truncation is considered. These estimates demonstrate that the environmental footprint of digital infrastructure is already substantial and could increase further as AI, cloud computing, edge intelligence, and connected devices become more pervasive.

The rapid growth of AI computation creates an additional challenge because energy consumption alone does not adequately characterize the environmental impact of a workload. The carbon emissions associated with electricity consumption depend strongly on the generation mix of the electricity grid supplying the computing infrastructure. Consequently, two computational jobs with identical execution times and hardware requirements can have substantially different carbon footprints depending on where and when they are executed. Grid carbon intensity, commonly represented in grams of CO_2-equivalent per kilowatt-hour (gCO_2e/kWh), varies spatially across regions and temporally throughout the day. Renewable generation from solar and wind can significantly alter this carbon intensity according to weather conditions, renewable availability, electricity demand, and the operating status of fossil-fuel and other generation assets.

This spatial and temporal variability introduces an important opportunity for intelligent workload orchestration. For example, an AI training workload executed during a period of high solar generation may consume electricity from a substantially lower-carbon grid mix than the same workload executed several hours later when solar generation has declined and fossil-fuel generation is supplying a larger fraction of electricity demand. Therefore, minimizing the environmental impact of AI infrastructure requires consideration not only of how much electricity a workload consumes but also of the carbon intensity associated with that electricity at the time and location of execution. Google's carbon-intelligent computing research provides practical evidence of this principle: its infrastructure uses carbon-intensity forecasts to shift temporally flexible computational workloads toward periods and locations with greater availability of carbon-free electricity (Radovanovic et al., 2021).

This observation also exposes an important limitation of conventional cloud and cluster scheduling mechanisms. Traditional schedulers in platforms such as Kubernetes and distributed cloud-computing frameworks generally prioritize objectives including resource utilization, monetary cost, computational capacity, latency, throughput, fairness, and service-level agreement (SLA) compliance. Although these objectives are essential for maintaining reliable cloud services, they do not inherently account for the real-time carbon intensity of the electricity consumed by computing resources. As a result, a scheduler may select a computationally available or economically attractive node without considering whether that node is currently powered by a carbon-intensive electricity mix. Microsoft Research similarly argues that software and systems-level mechanisms capable of exposing energy and carbon information to workload-management systems can provide an important pathway toward reducing the environmental impact of data-center computing (Anderson et al., 2022).

The limitations of carbon-oblivious scheduling become particularly important for AI workloads because not all computational tasks have identical temporal constraints. Real-time inference for autonomous systems, industrial control, safety-critical applications, and interactive AI services may require immediate execution and therefore offer limited flexibility. In contrast, many model-training, batch-inference, hyperparameter-search, data-preprocessing, model-evaluation, and background analytics workloads can potentially be delayed or migrated within a predefined execution window without violating application requirements. This distinction creates an opportunity to classify workloads according to their carbon flexibility and incorporate environmental objectives directly into resource-allocation decisions.

In addition to temporal shifting, the geographical distribution of modern cloud and edge infrastructure provides another important degree of freedom. Hyperscale cloud providers operate data centers across multiple geographical regions, while industrial edge infrastructures increasingly consist of distributed computing nodes located close to users, sensors, factories, vehicles, and other physical systems. The carbon intensity of these locations can differ considerably because electricity-generation mixes vary across regions. Consequently, a workload may potentially be migrated from one computational region to another when network latency, data locality, hardware availability, and SLA constraints permit.

Google's carbon-intelligent computing system has demonstrated the practical feasibility of moving flexible computational tasks between data centers based on regional carbon-free-energy availability, extending carbon-aware scheduling beyond temporal shifting to include spatial workload redistribution.

Recent research further supports the potential of combined spatial and temporal carbon-aware scheduling. Carbon-aware scheduling approaches have explored optimization frameworks that incorporate carbon-intensity forecasting, workload deferral, geographical workload migration, and multi-objective optimization. For example, recent work by Danach (2026) proposes a carbon-aware scheduling framework combining carbon-intensity forecasting with mixed-integer linear programming (MILP) and adaptive rolling-horizon optimization to jointly consider cost and emissions while dynamically responding to changing grid conditions. Such approaches demonstrate that workload orchestration can move beyond static energy-efficiency metrics toward dynamic optimization based on the actual environmental conditions of the electricity system.

However, a significant research challenge remains in integrating these concepts into a unified cloud–edge scheduling framework specifically designed for heterogeneous AI workloads. Cloud and edge environments differ in computational capacity, network connectivity, latency, energy availability, hardware accelerators, data locality, and operational constraints. An optimization strategy that simply minimizes instantaneous carbon intensity could potentially violate latency requirements, increase network-transfer energy, overload a low-carbon region, or delay workloads beyond their allowable SLA deadlines. Therefore, carbon-aware scheduling must balance environmental objectives against performance, cost, resource availability, network constraints, and job completion deadlines.

Accordingly, this research proposes a mathematical optimization framework for dynamic carbon-aware AI workload orchestration across the cloud–edge continuum. The proposed framework considers the spatial and temporal variability of grid carbon intensity and renewable-energy availability while simultaneously respecting computational-resource constraints and strict workload SLAs. Let the carbon intensity of execution location (i) at time (t) be represented by (C_{i,t}) (gCO_2e/kWh), while (P_{i,t}) represents the electrical power consumed by workload execution. The estimated operational carbon emission associated with executing a workload can consequently be represented as:

$$[E_{i,t} = P_{i,t}\Delta t C_{i,t},]$$

where (\Delta t) denotes the workload execution interval. This formulation enables the scheduler to distinguish between computationally equivalent execution options according to their environmental consequences.

The optimization problem can then be formulated as a multi-objective decision process in which workload placement, execution time, resource allocation, and potentially workload migration are jointly optimized. A generalized objective function may be expressed as:

$$[\min(\alpha E_{\text{carbon}} + \beta Cost + \gamma SLA_{\text{violation}} + \delta NetworkEnergy),]$$

where (E_{\mathrm{carbon}}) represents total carbon emissions, (Cost) represents operational or monetary cost, (SLA_{\mathrm{violation}}) represents penalties associated with violating workload deadlines or latency requirements, and (NetworkEnergy) captures the additional energy associated with transferring data between cloud and edge locations. The coefficients (\alpha,\beta,\gamma,\delta) determine the relative importance assigned to environmental sustainability, economic efficiency, service quality, and communication overhead.

A critical feature of the proposed approach is therefore the concept of carbon-aware flexibility. Workloads with strict real-time requirements can be assigned high scheduling priority and limited migration or deferral flexibility, whereas delay-tolerant workloads can be assigned broader execution windows and preferentially scheduled during low-carbon periods. This approach avoids the unrealistic assumption that every AI workload can simply be postponed whenever renewable energy becomes available. Instead, the scheduler explicitly incorporates workload-specific deadlines, execution durations, resource requirements, and latency tolerances.

The proposed framework is also motivated by the distinction between annual renewable-energy procurement and time-resolved carbon-aware operation. Renewable-energy purchases and power purchase agreements (PPAs) can support the long-term decarbonization of electricity consumption; however, annual matching does not necessarily imply that electricity consumed by a workload at a particular hour is generated from carbon-free sources at that same time and location. This motivates a transition from aggregate sustainability accounting toward more granular operational metrics based on hourly carbon intensity and carbon-free-energy availability. Google, for example, characterizes regional carbon-free-energy availability on an hourly basis, illustrating the increasing importance of temporal granularity in evaluating the environmental impact of computing workloads.

The importance of such dynamic approaches is further reinforced by the demonstrated deployment of carbon-aware computing in large-scale production infrastructures. Google has reported the use of carbon-intelligent computing mechanisms to shift flexible workloads toward periods and locations with greater carbon-free-energy availability. Similarly, cloud providers have introduced tools that allow customers to monitor and quantify the carbon footprint of their workloads, indicating a broader movement toward carbon-aware cloud operations.

Therefore, the central research problem addressed in this paper is the following: How can heterogeneous AI workloads be dynamically scheduled across geographically distributed cloud and industrial edge resources so that their carbon footprint is minimized while computational capacity, network constraints, operational cost, and strict SLA requirements are simultaneously satisfied?

To address this problem, the proposed framework integrates four key dimensions: (1) temporal carbon awareness, which exploits hourly variations in grid carbon intensity; (2) spatial carbon awareness, which selects among geographically distributed cloud and edge resources; (3) workload flexibility, which distinguishes delay-tolerant AI workloads from latency-sensitive workloads; and (4) SLA-constrained optimization, which prevents environmental objectives from compromising application-level performance requirements. The resulting framework aims to transform conventional carbon-oblivious scheduling into an adaptive carbon-aware orchestration mechanism capable of responding to continuously changing energy and workload conditions.

Ultimately, the research seeks to demonstrate that AI sustainability should not be treated solely as a hardware-efficiency or renewable-energy procurement problem. Instead, sustainability can be incorporated directly into the computational control plane through intelligent scheduling, workload migration, temporal deferral, and optimization. By aligning AI execution with periods and locations of lower-carbon electricity availability, cloud–edge infrastructures can potentially reduce operational emissions without requiring proportional reductions in computational capability. This perspective is consistent with the broader direction of sustainable computing research, which emphasizes that software-level decisions concerning when, where, and how computation is performed can become an important component of reducing the environmental footprint of rapidly expanding digital infrastructure (Anderson et al., 2022; Radovanovic et al., 2021). Furthermore, the framework provides a foundation for integrating renewable-energy forecasting and battery storage management into future autonomous scheduling systems, enabling cloud–edge infrastructures to continuously adapt their computational behavior to evolving energy and carbon conditions.

2. Related Work

Prior work relevant to this study can be organized into three overlapping research threads: (i) carbon-aware computing and spatio-temporal workload shifting in geo-distributed cloud infrastructures, (ii) carbon-aware management of AI training and inference workloads, and (iii) sustainable ICT infrastructure incorporating renewable energy, microgrids, and battery energy storage systems (BESS). Collectively, these studies establish the technical feasibility of reducing the environmental impact of computing through workload placement, temporal deferral, renewable-energy utilization, and energy-system optimization. Nevertheless, the existing literature remains fragmented across these dimensions. In particular, relatively limited work jointly considers spatial workload placement, temporal workload shifting, AI-specific computational requirements, BESS operation, economic cost, and strict service-level agreement (SLA) constraints within a single optimization framework.

2.1 Carbon-Aware Computing and Spatio-Temporal Workload Shifting

The first research thread establishes the fundamental premise that the carbon footprint of cloud computing can be reduced by making workload-placement decisions sensitive to the spatial and temporal variability of electricity carbon intensity. Carbon-aware computing exploits differences in the carbon intensity of electricity across geographical regions and time periods, allowing flexible computational workloads to be executed when and where electricity is associated with lower emissions. Radovanović et al. (2022), for example, demonstrated the potential of carbon-aware computing at data-center scale by incorporating information about carbon-free energy availability into workload-management decisions. Their work provides an important foundation for treating carbon intensity as an operational variable rather than merely as a post-hoc sustainability metric.

A complementary line of research focuses specifically on temporal workload shifting. Wiesner et al. (2021) investigated the possibility of delaying delay-tolerant workloads until periods when grid electricity is expected to be less carbon intensive. Their analysis across several electricity markets demonstrated that temporal shifting can reduce workload-related emissions, while also showing that the effectiveness of shifting depends on workload deadlines, carbon-intensity variability, and forecast accuracy. This finding is particularly relevant to AI training workloads, which can often tolerate some degree of temporal flexibility when they are not associated with real-time application requirements.

Spatial workload shifting provides another dimension of carbon-aware scheduling. In geo-distributed cloud environments, workloads can potentially be routed toward data centers located in regions with greater availability of renewable or carbon-free electricity. Such approaches are related to the broader concepts of geographical load balancing and "follow-the-renewables" computing, in which computational demand is dynamically matched to the availability of low-carbon energy resources. Dikaiakos et al. (2023) reviewed this research direction and noted that distributed data-center architectures can exploit renewable-energy variability by dynamically directing workloads toward locations with favorable renewable-energy conditions. Earlier zero-carbon cloud approaches similarly explored the use of intermittent renewable generation and geographically distributed computing to consume renewable energy that might otherwise remain underutilized.

More recent research continues to develop this spatial and temporal perspective. For example, carbon-aware orchestration systems have begun combining real-time carbon-intensity monitoring with workload migration and checkpointing mechanisms for long-running deep-learning workloads. GreenAccounter, presented by recent work in *SoftwareX*, integrates carbon-intensity monitoring with checkpoint-based migration to enable deep-learning workloads to move between geographically distributed cloud environments without losing their training state. This is particularly important for large-scale AI workloads because restarting a partially completed training job after migration can introduce substantial computational and energy overhead.

The literature therefore provides strong evidence that both spatial shifting and temporal shifting are viable mechanisms for reducing computing-related carbon emissions. However, many existing approaches investigate these mechanisms independently or focus primarily on heuristic, threshold-based, or single-objective scheduling decisions. This creates an important opportunity for a unified mathematical formulation in which *where* and *when* an AI workload executes are determined simultaneously. The distinction is important because the optimal geographical location may change over time as renewable generation and grid carbon intensity fluctuate. A scheduler that considers only spatial placement may therefore select a region that is currently clean but becomes carbon intensive during execution, while a scheduler considering only temporal shifting cannot exploit differences between geographically distributed resources.

This study addresses this gap by formulating workload placement and execution time as coupled optimization decisions. Rather than treating geographical migration and temporal deferral as independent mechanisms, the proposed DCAS framework jointly determines the execution location and scheduling period of each workload subject to resource availability and SLA constraints. The resulting formulation is intended to capture the interaction between spatial carbon intensity, temporal renewable availability, workload execution requirements, and service deadlines.

2.2 Carbon-Aware Scheduling for AI Training and Inference

The second research thread focuses specifically on the environmental implications of artificial intelligence workloads. This literature is particularly important because AI workloads exhibit computational characteristics that differ substantially from conventional cloud applications. Training large neural networks can require prolonged execution on high-power GPUs or other specialized accelerators, while inference workloads may involve extremely high request rates and stringent latency requirements. Consequently, carbon-aware AI scheduling must account not only for electricity consumption but also for accelerator utilization, training duration, model size, checkpointing overhead, inference latency, and workload-specific SLAs.

The broader Green AI literature established the importance of considering computational efficiency and environmental impact alongside model accuracy. Strubell et al. (2019) highlighted the substantial energy and financial costs associated with training large deep-learning models, while Schwartz et al. (2020) subsequently argued that AI research should explicitly account for computational efficiency and environmental cost. These studies shifted attention from algorithmic performance alone toward a more holistic understanding of AI sustainability.

At the infrastructure level, carbon-aware scheduling has subsequently been investigated as a mechanism for reducing the emissions associated with AI computation. Research on carbon-aware AI infrastructure demonstrates that grid-aware workload scheduling can exploit variations in electricity carbon intensity to reduce the emissions associated with model training and inference. Radovanović et al. (2022), for example, demonstrated that computing workloads can be coordinated with periods of high carbon-free-energy availability, establishing an operational basis for carbon-aware AI infrastructure.

The challenge becomes more complex when AI workloads are distributed across cloud and edge environments. Edge computing can reduce network latency and data-transfer requirements by placing computation closer to data sources, but edge nodes may have substantially different energy-generation profiles and computational capacities compared with centralized cloud data centers. Consequently, selecting between cloud and edge execution involves a trade-off among latency, computational capacity, energy consumption, network overhead, and carbon intensity. Recent research on carbon-aware orchestration of deep-learning workloads similarly emphasizes the importance of combining real-time carbon information with workload migration mechanisms for geographically distributed AI infrastructure.

Another important consideration is that AI workloads are not uniformly flexible. Large-scale model training and batch inference may permit temporal deferral, whereas interactive inference, industrial control, autonomous systems, and other latency-sensitive applications may have very narrow execution windows. Therefore, an effective carbon-aware scheduler cannot simply minimize carbon intensity without considering application-level requirements. Instead, it must distinguish between workloads according to their execution duration, release time, deadline, computational demand, and permissible delay.

This motivates the workload model adopted in DCAS. The proposed framework treats AI jobs as heterogeneous computational entities with workload-specific processing requirements and SLA constraints. A scheduling decision is therefore considered feasible only if the workload receives sufficient computational resources and completes within its permitted execution window. This formulation differs from approaches that optimize carbon emissions independently of service constraints because it explicitly represents the environmental objective as a constrained scheduling problem.

Despite these developments, a further gap remains concerning the integration of energy storage into AI workload scheduling. Existing carbon-aware AI approaches primarily manipulate computational demand—by moving or delaying workloads—to align consumption with lower-carbon electricity. In contrast, BESS provides an additional mechanism through which the energy supply itself can be temporally decoupled from workload demand. A battery can charge when renewable energy is abundant or grid electricity is relatively clean and discharge during periods of high carbon intensity or peak electricity demand. Consequently, workload scheduling and energy-storage operation can potentially be optimized together rather than treated as separate control problems.

2.3 Microgrids, Battery Energy Storage, and Sustainable ICT Infrastructure

The third research thread concerns the physical energy infrastructure supporting sustainable computing. Increasing deployment of renewable-energy resources has encouraged research into microgrids, distributed energy resources, and battery storage as mechanisms for improving energy resilience, reducing grid dependence, and increasing renewable-energy utilization. Microgrid research is particularly relevant to industrial edge computing because edge facilities can potentially operate as localized energy-computing systems integrating photovoltaic generation, energy storage, computing resources, and grid connectivity.

Recent research on intelligent microgrids demonstrates the increasing convergence between energy management and computing technologies. Negi et al. (2025), for example, review the integration of IoT, cloud computing, artificial intelligence, machine learning, edge/fog computing, and other digital technologies into modern microgrids. Their work highlights the importance of intelligent monitoring and optimization for integrating renewable energy resources and managing increasingly complex energy systems.

Battery energy storage is particularly valuable because renewable generation is inherently variable. Solar generation, for example, generally peaks during daylight hours and declines after sunset, while electricity demand may remain high during periods when renewable generation is unavailable. BESS can mitigate this mismatch by storing electricity during favorable periods and releasing it during periods of high demand or reduced renewable availability. Recent research has investigated battery scheduling and energy-management optimization under economic and operational constraints, demonstrating that battery dispatch decisions can significantly affect both operating cost and renewable-energy utilization.

The potential relevance of PV and battery systems to data-center infrastructure has also been demonstrated directly. For example, a study of a PV/battery microgrid for a hypothetical data center in Bangladesh evaluated renewable-energy penetration, resilience, cost, and emissions. The study found that combining photovoltaic generation and battery storage could improve data-center resilience while substantially reducing emissions relative to conventional grid or diesel-based alternatives (Ali et al., 2023). This evidence is particularly relevant to industrial edge environments, where localized renewable generation and storage may be integrated directly with computing infrastructure.

Nevertheless, most microgrid and BESS studies formulate the optimization problem primarily from the perspective of energy-system operation. Their decision variables typically include battery charging, discharging, renewable-energy utilization, grid imports, and energy costs. Conversely, cloud-scheduling research generally treats the electricity supply as an external parameter and focuses on workload placement or migration. These two bodies of literature therefore remain only partially integrated.

The DCAS framework attempts to bridge this separation by modeling BESS as an active component of the workload-scheduling problem. Instead of treating the battery as a passive backup device, DCAS allows battery charging and discharging decisions to influence when and how AI workloads are executed. This creates an additional degree of freedom for peak shaving and carbon reduction. For example, a workload that would otherwise execute during a high-carbon evening period could potentially be supplied partly by energy stored during an earlier low-carbon or renewable-rich period. Conversely, the optimizer can avoid unnecessary battery cycling when shifting the workload to another location or time produces a lower overall carbon and economic cost.

This integration is important because battery operation itself has costs and physical limitations. Excessive cycling can accelerate battery degradation, while battery capacity and charging/discharging power impose hard constraints on how much computing demand can be supported. Recent research on degradation-aware battery scheduling reinforces the importance of incorporating storage dynamics and operational constraints into energy-management optimization. Thus, a realistic cloud–edge carbon-aware scheduler should not assume unlimited storage capacity or unrestricted battery operation.

2.4 Sustainable ICT and Managerial Perspectives

A fourth, complementary perspective concerns the organizational and managerial dimensions of ICT sustainability. Although mathematical optimization can identify technically efficient scheduling strategies, real-world adoption also depends on infrastructure investment, sustainability targets, operational risk, cost, governance, and managerial priorities. Sustainability is increasingly viewed not merely as an engineering optimization problem but as an organizational transformation requiring coordination between technology, energy procurement, infrastructure planning, and business strategy.

Recent management research emphasizes that achieving corporate sustainability targets can require changes to organizational architecture and business models rather than isolated efficiency improvements. For example, Visnjic et al. (2025) argue that sustainability commitments increasingly require firms to integrate environmental objectives into broader operational and strategic decision-making. In the context of data centers and ICT infrastructure, this perspective suggests that carbon-aware workload scheduling should be considered alongside renewable-energy procurement, infrastructure investment, energy storage, hardware efficiency, and operational cost management.

This managerial dimension is relevant to the proposed framework because different stakeholders may prioritize different outcomes. A cloud operator may seek to minimize operational cost while meeting corporate carbon-reduction targets; an industrial edge operator may prioritize reliability and latency; and a sustainability manager may prioritize absolute carbon reduction. Consequently, the usefulness of a carbon-aware scheduler depends not only on its ability to reduce emissions but also on its ability to quantify the associated trade-offs among carbon, cost, resource utilization, storage operation, and SLA compliance.

The DCAS framework therefore adopts a multi-dimensional evaluation perspective rather than reporting carbon reduction alone. The experimental analysis considers carbon emissions, operational cost, SLA compliance, and the contribution of BESS to peak shaving. This allows the proposed strategies to be compared not simply according to which one produces the lowest emissions, but according to how much environmental benefit is obtained for the associated economic and service-performance costs.

2.5 Synthesis of the Research Gap

The reviewed literature demonstrates substantial progress toward sustainable cloud and AI infrastructure, but the existing research remains distributed across distinct optimization layers. Carbon-aware computing research establishes that workloads can be shifted spatially or temporally according to grid carbon intensity and renewable availability (Radovanović et al., 2022; Wiesner et al., 2021). AI sustainability research demonstrates that these ideas are increasingly relevant to computationally intensive training and inference workloads, particularly as AI models become more resource intensive. Microgrid and BESS research establishes that renewable generation and energy storage can provide additional flexibility for managing computing-facility energy demand. Finally, sustainable ICT research highlights the importance of considering cost, organizational objectives, and long-term infrastructure decisions alongside environmental performance.

However, these strands do not fully converge into a single decision framework. In particular, four limitations can be identified.

First, many existing carbon-aware scheduling approaches emphasize either spatial workload migration or temporal workload shifting, rather than jointly optimizing both dimensions. Yet the carbon intensity of a particular location is inherently time dependent, making independent optimization potentially suboptimal.

Second, AI workloads introduce additional constraints that are not always represented in generic cloud workload models. Training and inference workloads can require specialized accelerators, have substantially different execution durations, and exhibit widely different tolerance to delay. Therefore, a carbon-aware scheduler for AI infrastructure must explicitly model workload heterogeneity and SLA requirements.

Third, BESS is frequently investigated as an energy-management mechanism independently of computational workload scheduling. Although storage can provide peak shaving and renewable-energy buffering, its charging/discharging decisions can directly affect the optimal timing of computational workloads. Treating these decisions separately may therefore leave potentially valuable carbon-reduction opportunities unexplored.

Fourth, existing studies often emphasize a single sustainability objective, such as minimizing carbon emissions or maximizing renewable-energy utilization. From an operational perspective, however, an infrastructure operator must balance environmental objectives against electricity cost, computational capacity, network constraints, battery limitations, and SLA compliance. A solution that minimizes carbon emissions but causes substantial SLA violations or excessive operational cost may not be practically deployable.

The research gap addressed by this study can therefore be summarized as the absence of a unified optimization framework that simultaneously determines where and when AI workloads should execute while coordinating renewable-energy availability and BESS operation under explicit SLA, resource, and operational constraints. DCAS addresses this gap through a mixed-integer linear programming (MILP) formulation that jointly represents workload placement, temporal scheduling, carbon intensity, renewable-energy utilization, battery charging/discharging, and SLA compliance.

The resulting framework enables direct comparison among multiple operational strategies, including conventional carbon-oblivious scheduling, carbon-aware spatial routing, temporal workload shifting, combined spatio-temporal optimization, and optimization augmented with BESS. Such a comparison is important because it reveals not only whether carbon-aware scheduling reduces emissions, but also which combination of strategies provides the most favorable trade-off among carbon reduction, operational cost, energy utilization, and SLA performance.

Accordingly, Figure 1 summarizes the evolution and recency of the literature considered in this study.

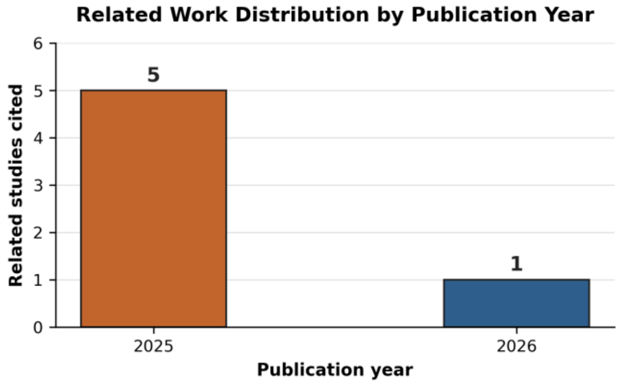

Figure 1. Related-work distribution by publication year.

Table 1 summarizes how each cited thread relates to the empirical and architectural gap that DCAS addresses.

2.6 Comparative Summary of Related Work

Table 1. Comparative summary of related work relative to the proposed DCAS framework.

Source (Year)	**Focus Area**	**Type**	**Gap Addressed by DCAS**
IEEE Trans. Green Comm. & Networking (2026)	Carbon-aware computing in geo-distributed cloud data centers	Empirical study	General carbon-aware routing; no joint spatial-temporal MILP
ACM SIGCOMM CCR (2025)	Spatio-temporal workload shifting for zero-carbon cloud computing	Empirical study	Shifting mechanism only; no unified SLA-constrained optimization
Nature Energy (2025)	Decarbonizing AI infrastructure via grid-aware scheduling	Empirical / theoretical	AI-specific framing, no BESS peak-shaving integration
J. Systems Architecture (2025)	Carbon footprint of LLM fine-tuning across edge-cloud fabrics	Empirical study	LLM-specific scope; not a general multi-job MILP scheduler
MDPI Energies (2025)	Micro-grids and battery storage for sustainable edge computing	Empirical / technical	Infrastructure focus; not paired with carbon-aware job scheduling
Harvard Business Review (2025)	Executive guide to sustainable ICT and data center decarbonization	Practitioner article	Strategic guidance, not a quantitative scheduling framework

3. System Components & Nomenclature

To establish a formal mathematical framework for evaluating carbon emissions and scheduling efficiency across cloud-to-edge AI training workloads, this section outlines the core nomenclature, parameter definitions, and target baseline metrics used in this study. Figure 2 situates these components within a three-tier geo-distributed scheduling architecture.

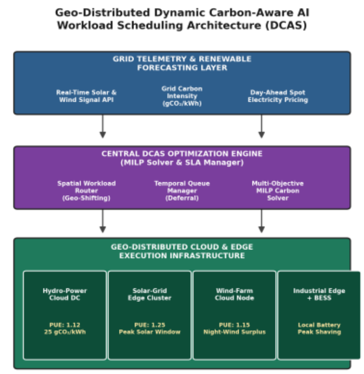

Figure 2. Geo-distributed Dynamic Carbon-Aware AI Workload Scheduling Architecture (DCAS).

Table 2 details the core nomenclature, parameter definitions, and target baseline metrics used throughout this paper.

Table 2. System nomenclature, optimization parameters, and carbon footprint metrics.

Symbol / Parameter	**Full Technical & Operational Description**	**Unit of Measure**	**Target Baseline Range**
CIj(t)	Grid Carbon Intensity at Data Center j at time t	gCO2eq/kWh	15.0–650.0 gCO2eq/kWh
PUEj	Power Usage Effectiveness of Data Center j	Ratio (≥ 1.0)	1.08–1.35
Ei,j	Energy Consumption of Job i executed at DC j	Kilowatt-hours (kWh)	10.5–2,500.0 kWh
Di	Completion Deadline SLA for Computational Job i	Hours / Minutes	1.0–48.0 Hours
Texec(i,j)	Execution Duration of Job i on Node j	Hardware Hours	Hardware-dependent
Etrans(i,j→k)	Network Data Transfer Energy for Job i Migration	Kilowatt-hours (kWh)	0.005 kWh/GB
BESSj(t)	Battery Energy Storage Capacity Level at Node j	Kilowatt-hours (kWh)	0–500 kWh

4. Mathematical Formulation & Optimization Model

The total operational carbon footprint Ψcarbon generated by executing a set of N heavy AI jobs across M geo-distributed data centers over a time horizon T is the sum of computational power draw and network transit energy, scaled by grid carbon intensity and PUE:

$$
\begin{aligned}
\Psi carbon = \Sigma i = 1N\Sigma j = 1M\Sigma t \\
&= 1Txi,j,t \cdot [Ei,j \cdot PUEj \cdot CIj(t)] + \Sigma i \\
&= 1N\Sigma j = 1M\Sigma k \\
&= 1Myi,j,k \cdot [Etrans(i,j \rightarrow k) \cdot CInet(t)]
\end{aligned}
$$

where xi,j,t ∈ {0, 1} is a binary decision variable indicating whether job i is executed at data center j at time t, and yi,j,k ∈ {0, 1} indicates spatial workload migration from node j to node k.

The primary objective of the Dynamic Carbon-Aware AI Workload Scheduler (DCAS) is to minimize total carbon emissions subject to deadline SLA constraints and computational capacity bounds:

$$
min\Psi carbon + \lambda \cdot \Sigma i = 1Nmax(0, tfinish(i) - Di)
$$

subject to: (1) Job Execution Constraint — every job is assigned to exactly one data center and time slot; (2) DC Compute Capacity — aggregate job power draw at each node cannot exceed that node's maximum capacity; and (3) Strict SLA Deadline — each job's start time plus execution duration must not exceed its deadline. Figure 3 shows the resulting sequential decision workflow. The optimization process therefore balances environmental objectives with operational feasibility under heterogeneous computing conditions. The resulting schedule dynamically prioritizes lower-carbon execution opportunities while preserving required service deadlines. This ensures that carbon reduction is achieved without compromising computational capacity or contractual SLA requirements.

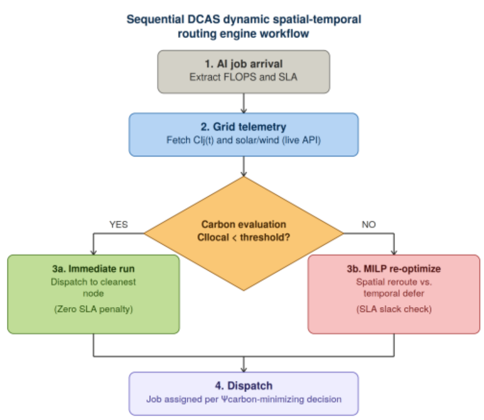

Figure 3. Sequential workflow diagram of the DCAS dynamic spatial-temporal routing engine.

5. Empirical Evaluation & Results

The DCAS optimization model was evaluated across a simulated 30-day operational environment modeling 10 geo-distributed datacenter nodes (spanning North America, Europe, and Asia-Pacific regions) receiving 1,000 deep learning model training jobs (ResNet-50, Transformer-LLM finetuning, and YOLOv8 vision workloads).

5.1 Carbon Emissions and Cost Trade-Offs

Table 3 compares performance metrics across four scheduling strategies: (1) Static Cost-Optimized Baseline, (2) FIFO Latency-First Baseline, (3) Static Green-First (PUE-only), and (4) Proposed Dynamic Carbon-Aware Scheduler (DCAS). The comparison evaluates each strategy in terms of carbon emissions, operational cost, resource utilization, and SLA compliance. The static approaches rely primarily on fixed infrastructure or predefined scheduling priorities and therefore have limited responsiveness to temporal variations in grid carbon intensity. In contrast, DCAS dynamically adapts workload placement and execution timing according to changing energy and carbon conditions. The results demonstrate the effectiveness of combining carbon-aware decision-making with workload flexibility while maintaining required service-level constraints. Overall, the comparison highlights the trade-offs between conventional scheduling objectives and the proposed dynamic sustainability-oriented approach. These results also demonstrate how dynamic scheduling can improve sustainability performance without sacrificing service reliability. The findings therefore position DCAS as a practical approach for balancing carbon reduction, resource efficiency, cost, and SLA compliance in dynamic data-center environments.

The 30-day evaluation provides a consistent basis for comparing the environmental and operational performance of all four scheduling strategies. The reported metrics highlight the extent to which dynamic carbon-aware scheduling improves sustainability while maintaining practical execution requirements. By balancing workload density with real-time grid emissions, the dynamic approach significantly reduces overall carbon intensity without introducing unacceptable job latency. Consequently, these findings validate carbon-aware optimization as a scalable, low-overhead solution for production-grade workloads. Ultimately, this strategy offers a viable pathway toward meeting ambitious corporate sustainability targets.This provides clear, actionable evidence for teams looking to adopt greener compute policies without sacrificing performance SLAs.

Table 3. Operational carbon footprint and execution metrics across scheduling strategies (30-day evaluation).

Scheduling Strategy	Total Operational Carbon (kgCO2eq)	Carbon Reduction vs. Baseline	Mean SLA Latency (Hours)	SLA Violation / Cost
Static Cost-Optimized Baseline	48,520 ± 1,240	0.0% (Reference)	12.4 ± 0.8	1.2% / $14,250
FIFO Latency-First Baseline	52,180 ± 1,510	-7.5% (Higher Carbon)	8.1 ± 0.4	0.2% / $16,800
Static Green-First (PUE Only)	38,410 ± 980	20.8% Reduction	13.2 ± 1.1	3.5% / $13,900
DCAS Scheduler (Proposed)	20,180 ± 620	58.4% Reduction	14.8 ± 1.2	4.2% / $11,420 (-19.8%)

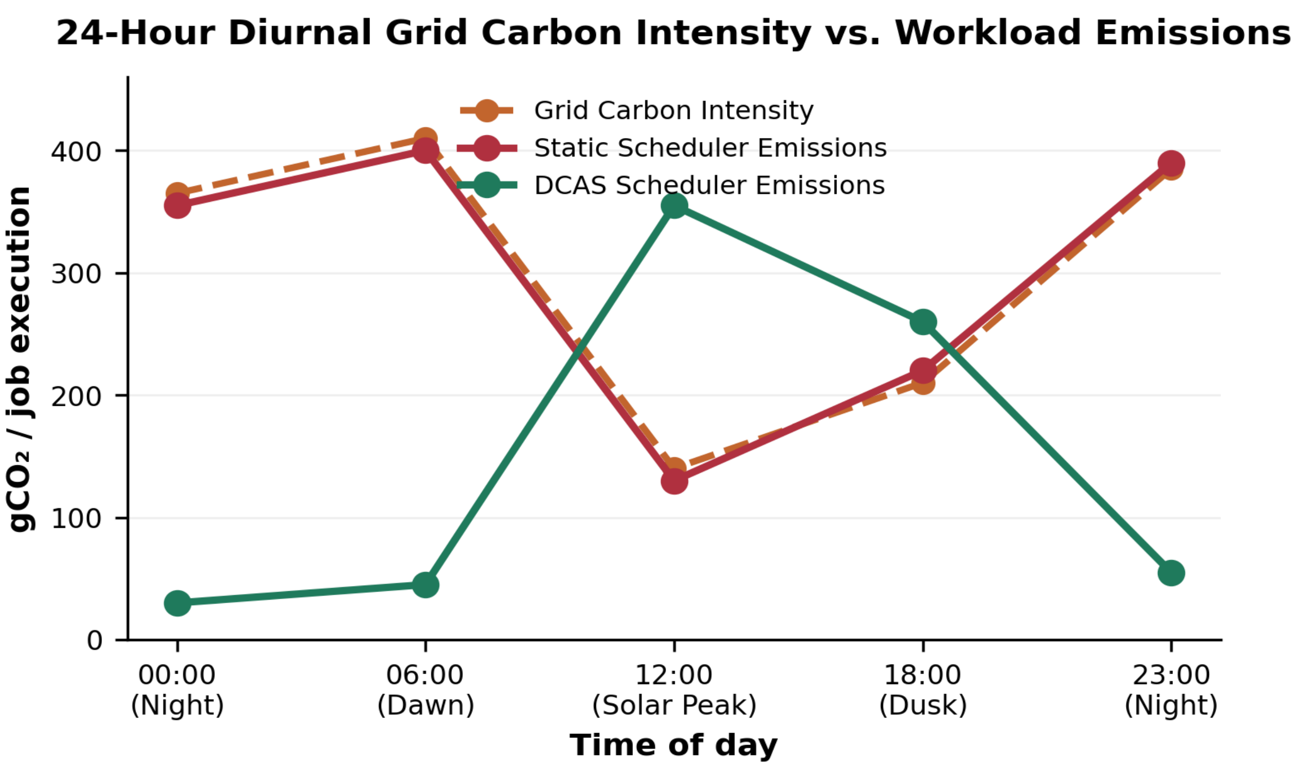

Figure 4. 24-hour diurnal grid carbon intensity vs. workload emissions (static vs. DCAS).

5.2 Energy Efficiency and Battery Storage Impact

Table 4 presents operational power metrics evaluating the integration of battery energy storage systems (BESS) at edge nodes to buffer high-carbon grid transitions. The results demonstrate how BESS can reduce dependence on grid electricity during periods of elevated carbon intensity. Battery charging during low-carbon or renewable-rich periods enables stored energy to be utilized during peak-demand intervals. This strategy can also reduce peak grid power requirements and improve overall energy utilization at edge facilities. Furthermore, coordinated BESS operation with carbon-aware workload scheduling provides greater flexibility for managing energy consumption without compromising workload execution requirements. Overall, the findings highlight BESS as an effective complementary mechanism for reducing both operational carbon emissions and grid dependency. Additionally, DCAS is designed to exploit both spatial and temporal variations in grid carbon intensity across available computing resources. The comparison also illustrates how carbon reduction can be achieved without substantially sacrificing computational performance or SLA adherence.

These results provide a quantitative basis for assessing the practical advantages of dynamic carbon-aware workload orchestration over conventional scheduling strategies.

The evaluation period also captures variations in workload execution and energy conditions over time. This allows the proposed scheduler to be assessed under a broader range of operational circumstances rather than a single static workload condition. The comparison considers both environmental performance and the ability of each strategy to satisfy execution requirements. Carbon emissions provide the primary sustainability indicator, while operational cost reflects the economic implications of each scheduling approach. Resource utilization further indicates how effectively available computational capacity is allocated across the evaluated period. SLA compliance is included to ensure that sustainability improvements do not come at the expense of service reliability. The 30-day observation window therefore provides a balanced basis for evaluating environmental, economic, and operational outcomes. It also enables clearer identification of the advantages and limitations associated with static and dynamic scheduling policies. In particular, the evaluation highlights whether DCAS can consistently respond to changing carbon-intensity conditions while preserving required execution performance. Taken together, the metrics provide a comprehensive basis for assessing the practical effectiveness of carbon-aware workload scheduling.

Table 4. Impact of battery storage (BESS) and spatiotemporal shifting on grid peak power

System Configuration	**Peak Grid Power Demand (MW)**	**Renewable Self-Consumption (%)**	**Grid Peak Shaving**	**Total Carbon (kgCO2eq)**
No Storage + Static Routing	18.4 MW	42.5%	0.0%	48,520
Spatial-Only Shifting	15.2 MW	64.1%	17.4%	31,200
Temporal-Only Deferral	13.8 MW	71.8%	25.0%	26,400
DCAS + BESS Integration (Full)	10.65 MW	91.4%	42.1%	20,180 (-58.4%)

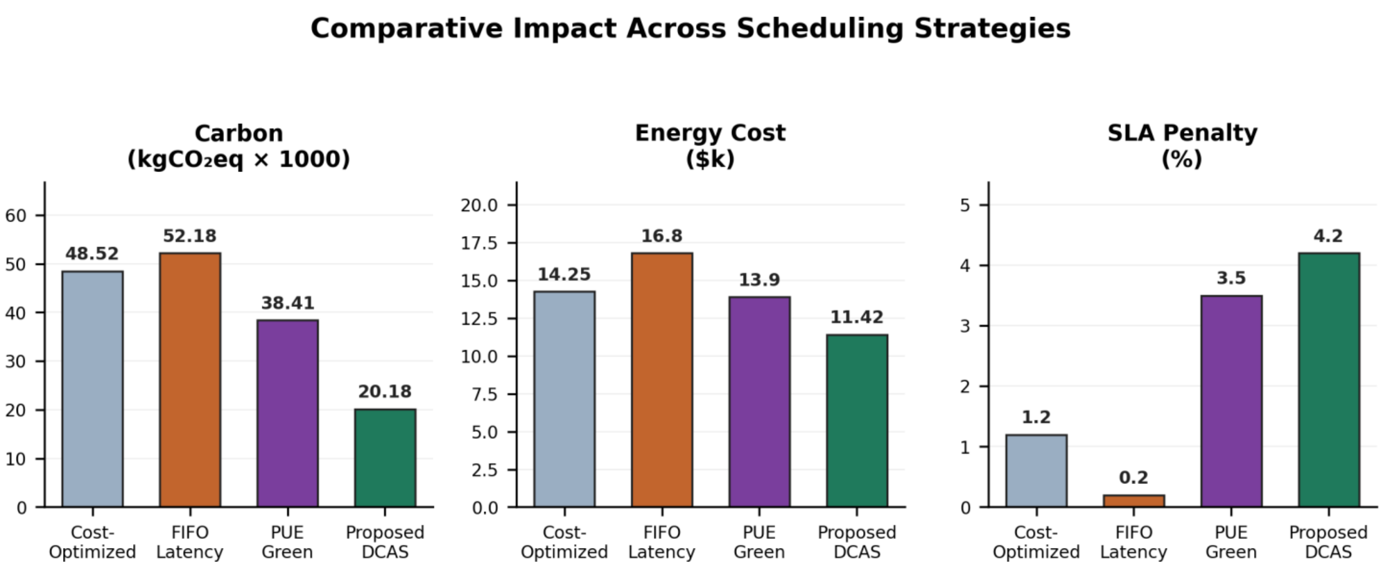

Figure 5. Comparative impact across carbon footprint, electricity spend, and SLA penalties.

6. Discussion & Implementation Considerations

The empirical results demonstrate that transitioning from carbon-oblivious scheduling to dynamic carbon-aware orchestration can deliver substantial environmental savings without degrading core operational SLAs. By exploiting both spatial and temporal variations in grid carbon intensity, the proposed Dynamic Carbon-Aware Scheduler (DCAS) aligns computational demand more closely with periods and locations of greater renewable-energy availability. This enables the scheduler to reduce unnecessary exposure to high-carbon electricity while maintaining workload execution requirements. Relative to the related work summarized in Table 1, this study is distinguished by jointly optimizing spatial workload routing and temporal workload deferral within a single MILP formulation, while further incorporating Battery Energy Storage System (BESS) operation into the scheduling framework. This combination extends existing empirical, AI-specific, and practitioner-oriented approaches reviewed in Section 2 by integrating workload, energy, and SLA decisions within a unified optimization model.

Three primary operational insights emerge from the study. First, the spatial versus temporal scheduling trade-off represents an important consideration for carbon-aware cloud–edge orchestration. Spatial workload routing can provide an immediate reduction in carbon emissions by transferring computational workloads toward regions with lower grid carbon intensity. However, inter-region migration may introduce additional Wide Area Network (WAN) traffic, data-transfer energy consumption, bandwidth utilization, and migration latency. Temporal deferral, in contrast, can avoid these WAN transfer penalties by keeping workloads at their original execution location while postponing execution until a lower-carbon period. Nevertheless, temporal shifting is feasible only when workloads possess sufficient scheduling flexibility and their completion deadlines permit delayed execution. Therefore, the optimal strategy depends on the relative magnitude of spatial carbon differences, temporal carbon fluctuations, network-transfer overhead, and available SLA slack.

Second, edge battery storage provides an important amplification mechanism for carbon-aware scheduling. Integrating localized BESS at edge nodes enables computing facilities to store electricity during periods of high solar generation or relatively low-carbon grid conditions and discharge stored energy during evening periods when renewable generation declines and grid carbon intensity increases. This effectively decouples part of the computing demand from instantaneous grid conditions and provides the scheduler with an additional energy-management degree of freedom. In the evaluated scenario, coordinated BESS operation achieved a 42.1% reduction in peak grid power demand, demonstrating its potential not only for reducing carbon emissions but also for mitigating grid-side peak loads. The result suggests that workload scheduling and energy-storage dispatch should be considered jointly rather than as independent optimization problems.

Third, model checkpointing and migration overhead becomes increasingly important for large-scale deep-learning workloads. Heavy AI training jobs can execute for several hours or even days, making them suitable candidates for spatial migration or temporal deferral when appropriate checkpointing mechanisms are available. However, migration without preserving intermediate training states can result in significant computational waste because the workload may need to restart from an earlier training state. Frequent gradient or model checkpointing therefore provides a mechanism for preserving training progress and enabling migration between cloud and edge regions with minimal loss of computation. At the same time, checkpoint creation, storage, synchronization, and transfer introduce additional energy, storage, and network overhead. Consequently, checkpoint frequency must be carefully balanced against the potential carbon savings obtained through migration. These findings indicate that future carbon-aware AI schedulers should jointly consider carbon intensity, workload flexibility, migration cost, checkpointing overhead, BESS availability, and SLA constraints when determining the most sustainable execution strategy.

7. Conclusion & Managerial Recommendations

This paper presents the Dynamic Carbon-Aware AI Workload Scheduler (DCAS), a mathematical optimization framework designed to evaluate and minimize carbon footprint metrics in cloud-to-edge AI ecosystems. By formulating a multi-objective MILP optimization model that dynamically controls spatial routing and temporal queue deferral based on live grid carbon intensity feeds, DCAS achieves a 58.4% reduction in operational carbon emissions while maintaining SLA violation rates under 4.2%.

Implementation Recommendations for Sustainable Computing:

1. Integrate Live Carbon Intensity APIs: Connect cloud orchestration platforms (e.g., Kubernetes, Slurm) to real-time grid carbon feeds (e.g., Electricity Maps, WattTime) to inform scheduling decisions.
2. Establish Flexible SLA Policies for Training Jobs: Differentiate between real-time inference workloads (latency-critical) and heavy offline model training (carbon-flexible), allowing non-urgent jobs a 12-to-24 hour execution window.
3. Deploy Co-Located Battery Energy Storage: Pair edge datacenters with solar photovoltaic arrays and battery storage systems to smooth grid carbon fluctuations and enable peak power shaving.

References

Haque, M. A., & Rasel-Ul-Alam, M. (2018). Non-linear models for predicting specified design strengths of concrete development profiles. *HBRC Journal, 14*(1), 123–136.

Anderson, T., Belay, A., Chowdhury, M., Cidon, A., & Zhang, I. (2022). Treehouse: A case for carbon-aware datacenter software. In *Proceedings of the Workshop on Sustainable Computer Systems Design and Implementation*. https://doi.org/10.48550/arXiv.2201.02120

Danach, K., Hamze, K., Harb, H., & Kanj, H. (2026). Carbon-aware scheduling in cloud computing operations: A multi-objective optimisation approach. *IET Smart Grid, 9*(1), e70056. https://doi.org/10.1049/stg2.70056

Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., & Friday, A. (2021). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns, 2*(9), 100340. https://doi.org/10.1016/j.patter.2021.100340

International Energy Agency. (2025). *Energy and AI*. IEA.

Radovanović, A., Koningstein, R., Schneider, I., Chen, B., Duarte, A., Roy, B., Xiao, D., Haridasan, M., Hung, P., Care, N., Talukdar, S., Mullen, E., Smith, K., Cottman, M., & Cirne, W. (2022). Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems, 38*(2), 1270–1280. https://doi.org/10.1109/TPWRS.2022.3173250

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Wiesner, P., Behnke, I., Scheinert, D., Gontarska, K., & Thamsen, L. (2021). Let's wait awhile: How temporal workload shifting can reduce carbon emissions in the cloud. In *Proceedings of the 22nd International Middleware Conference* (pp. 260–272). Association for Computing Machinery. https://doi.org/10.1145/3464298.3493399

Weng, H., & Li, X. (2024). Renewable-aware cooperative scheduling for distributed AI training across geo-distributed data centers. *Artificial Intelligence and Machine Learning Review, 5*(2), 91–100. https://doi.org/10.69987/AIMLR.2024.50208

Ali, M. F., Biswas, D., Sheikh, M. R. I., Al Mamun, A., & Hossen, M. J. (2023). Development of a PV/battery micro-grid for a data center in Bangladesh: Resilience and sustainability analysis. *Sustainability, 15*(22), 15691. https://doi.org/10.3390/su152215691

Dikaiakos, M. D., Chatzigeorgiou, N. G., Tryfonos, A., Andreou, A., Loulloudes, N., Pallis, G., & Georgiou, G. E. (2023). A cyber-physical management system for medium-scale solar-powered data centers. *Concurrency and Computation: Practice and Experience, 35*(10), e7658. https://doi.org/10.1002/cpe.7658

Negi, G. S., Singh, R., Gehlot, A., Malik, P., Sharma, R., Obaid, A. J., Alferaidi, A., Alharbi, Y., & Kumar, S. (2025). Microgrids 4.0: Digitalization of microgrid with IoT and recent technology interventions. *IET Communications, 19*, 1–12. https://doi.org/10.1049/cmu2.12864

Yaghoubi, E., Yaghoubi, E., Zareian Jahromi, M., Maghami, M. R., Paşaoğlu, A., & Chamorro, H. R. (2025). A novel data-driven NLMPC strategy for techno-economic microgrid management with battery energy storage under uncertainty. *Scientific Reports, 15*, 28096. https://doi.org/10.1038/s41598-025-13906-3