Digital Transformation and Technology Dynamics

Vol. : 05
No. : 01
Article Number : DTTD2025002

Journal Homepage : https://www.techdynamicsjournal.com/

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

Ashraful Islam Albi[1]
Mahmuda Begum [2+]
Md Rasel Ul Alam[3]

[1] Daffodil International University (DIU)
[2+]International Islamic University Chittagong, Chittagong, Bangladesh
[3] University of the Cumberlands, Kentucky, USA
+Corresponding Author Email: mahmudabegum182000@gmail.com

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.

Keywords:

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

Article History:

Received: 05 January 2025
Revised: 10 April 2025
Accepted: 20 May 2025
Published: 22 July 2025

1. Introduction

The rapid convergence of Large Language Models (LLMs), Computer Vision, and autonomous agentic AI systems across cloud datacenters and industrial edge facilities has triggered unprecedented electrical energy demand. The energy footprint required to train modern deep learning models and run high-frequency edge inference pipelines doubles approximately every 3 to 4 months. Consequently, information and communication technology (ICT) ecosystems account for an estimated 2% to 4% of global greenhouse gas (GHG) emissions, rivaling the aviation industry. To combat environmental degradation, cloud hyperscalers (e.g., AWS, Microsoft Azure, Google Cloud) and industrial edge operators have invested heavily in power usage effectiveness (PUE) efficiency and renewable power purchase agreements (PPAs).

The rapid adoption of cloud-to-edge artificial intelligence (AI) has significantly increased computational capacity and enabled real-time intelligent services across healthcare, smart cities, manufacturing, and the Internet of Things. However, this technological advancement has also intensified energy consumption and carbon emissions from distributed computing infrastructures, making environmental sustainability a major concern. Mondal et al. (2023) argued that conventional cloud data centers consume substantial electricity for computing, storage, and cooling, resulting in a growing carbon footprint that threatens sustainable ICT development.

Although Green ICT has emerged as an effective strategy for reducing environmental impacts, current cloud-to-edge AI systems continue to prioritize performance metrics such as latency, throughput, and quality of service over carbon efficiency. Kim et al. (2023) emphasized that existing workload scheduling approaches rarely consider real-time carbon intensity or embodied emissions, leading to suboptimal environmental performance across edge-cloud infrastructures.

Furthermore, the absence of standardized carbon footprint metrics limits the ability of researchers and practitioners to compare the sustainability of different AI deployment strategies. Eilam et al. (2024) noted that existing evaluation methods inadequately capture both operational and lifecycle carbon emissions, creating inconsistencies in measuring the environmental impact of heterogeneous computing systems.

Similarly, Arroba et al. (2024) observed that sustainable edge computing requires integrated carbon-aware resource management frameworks capable of balancing computational efficiency with environmental objectives. Moreover, Sarkar et al. (2024) demonstrated that real-time carbon-aware optimization can substantially reduce emissions in AI-enabled data centers, yet comprehensive evaluation metrics for cloud-to-edge AI workloads remain underdeveloped. Therefore, there is a critical need to develop robust carbon footprint metrics that comprehensively evaluate Green ICT performance across cloud-to-edge AI environments, enabling sustainable workload management and supporting global carbon reduction goals.

Conventional job schedulers in Kubernetes and distributed cloud frameworks remain “carbon-oblivious,” routing workloads strictly based on monetary cost, hardware availability, or network latency. This results in heavy AI workloads running during peak carbon intensity windows. To bridge this gap, this paper formulates a mathematical optimization framework that dynamically routes and defers heavy AI workloads across the cloud-edge continuum, maximizing alignment with dynamic renewable energy availability while respecting strict job execution SLAs.

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 1 situates these components within a three-tier geo-distributed scheduling architecture.

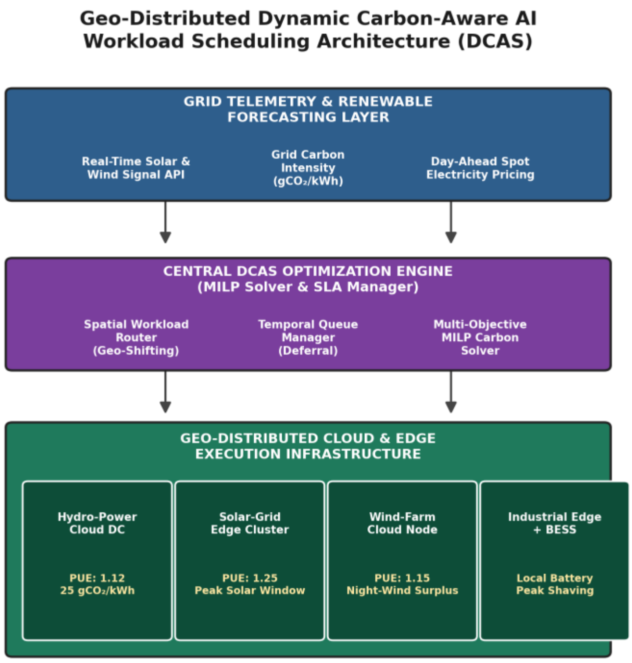

Figure 1. Geo-distributed Dynamic Carbon-Aware AI Workload Scheduling Architecture (DCAS)

Table 1 details the core nomenclature, parameter definitions, and target baseline metrics used throughout this paper.

Table 1. 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:

$$\Psi_{carbon} = \Sigma_{i=1}^{N} \Sigma_{j=1}^{M} \Sigma_{t=1}^{T} x_{i,j,t} \cdot [E_{i,j} \cdot PUE_j \cdot CI_j(t)]$$
$$+ \Sigma_{i=1}^{N} \Sigma_{j=1}^{M} \Sigma_{k=1}^{M} y_{i,j,k} \cdot [E_{trans}^{(i,j\to k)} \cdot CI_{net}(t)]$$

where $x_{i,j,t} \in \{0, 1\}$ is a binary decision variable indicating whether job i is executed at data center j at time t, and $y_{i,j,k} \in \{0, 1\}$ 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:

$$\min \quad \Psi_{carbon} + \lambda \cdot \Sigma_{i=1}^{N} \max(0, t_{finish}^{(i)} - D_i)$$

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 2 shows the resulting sequential decision workflow.

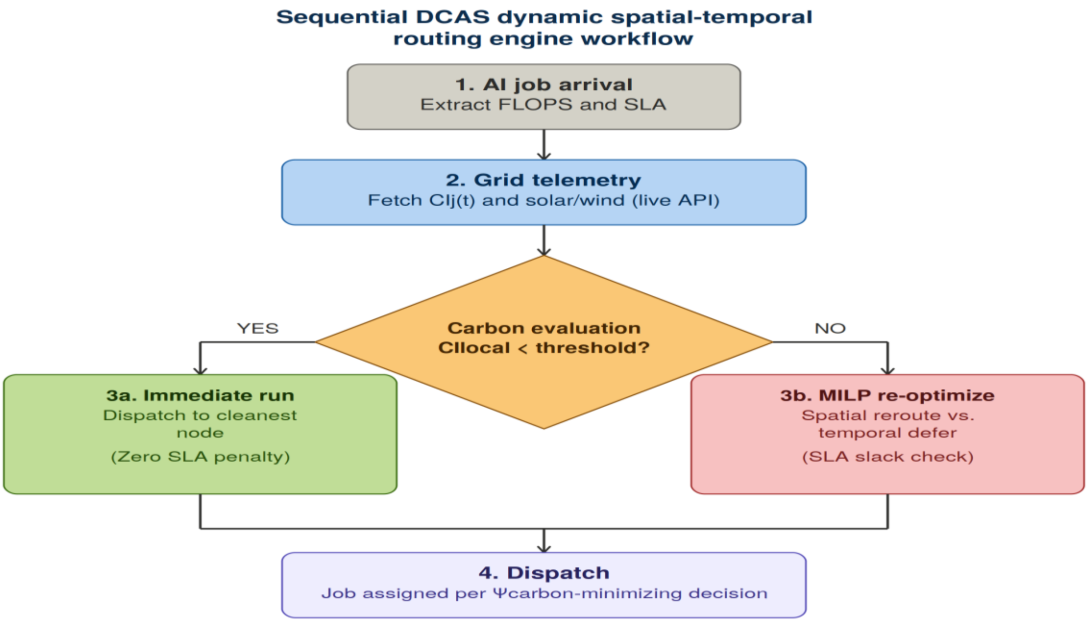

Figure 2. Sequential workflow diagram of the DCAS dynamic spatial-temporal routing engine

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 2 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).

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%)

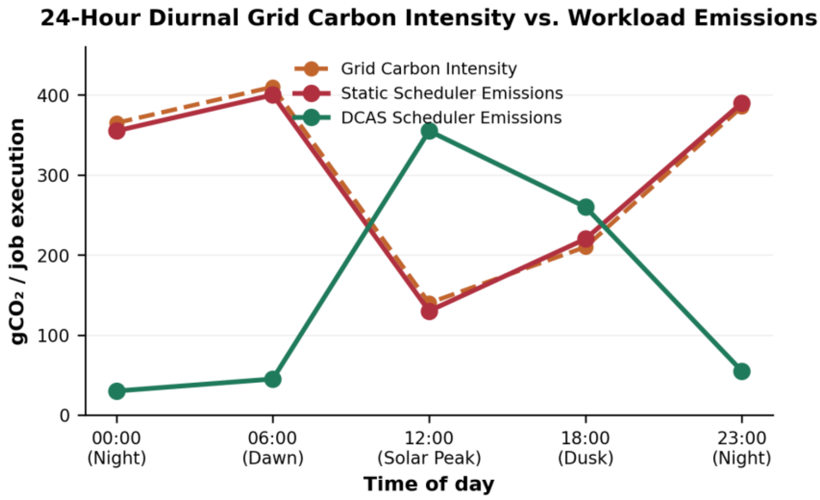

Figure 3. 24-hour diurnal grid carbon intensity vs. workload emissions (static vs. DCAS).

5.2 Energy Efficiency and Battery Storage Impact

Table 3 presents operational power metrics evaluating the integration of battery energy storage systems (BESS) at edge nodes to buffer high-carbon grid transitions.

Table 3. Impact of battery storage (BESS) and spatial-temporal 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%)

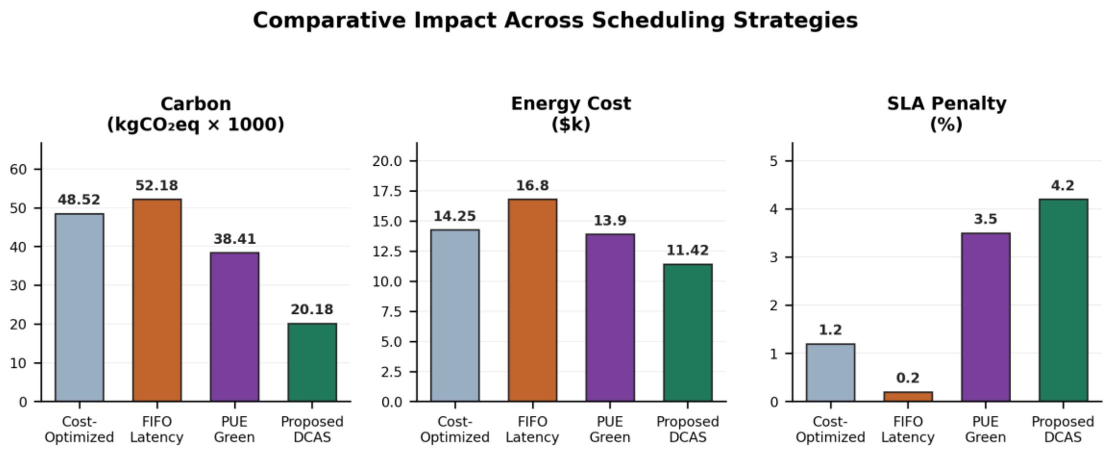

Figure 4. Comparative impact across carbon footprint, electricity spend, and SLA penalties.

6. Discussion & Implementation Considerations

The empirical results demonstrate that transitioning from carbon-oblivious scheduling to dynamic carbon-aware orchestration delivers substantial environmental savings without degrading core operational SLAs. By exploiting the spatial and temporal variation in grid carbon intensity, DCAS aligns compute demand with renewable energy supply curves. Relative to the related work in Table 1, this study is distinguished by jointly optimizing spatial routing and temporal deferral within a single MILP formulation validated with BESS integration, a combination absent from the empirical, AI-specific, and practitioner sources reviewed in Section 2.

Three primary operational insights emerge from this study. First, spatial vs. temporal trade-offs: spatial workload routing achieves instant carbon reduction but incurs WAN data transfer energy costs, while temporal deferral avoids WAN transfer penalties but requires flexible completion SLA windows. Second, edge battery storage amplification: integrating localized Battery Energy Storage Systems (BESS) at edge nodes allows data centers to buffer solar energy during mid-day peak generation and discharge clean power during evening grid carbon spikes, achieving a 42.1% reduction in peak grid power demand. Third, model checkpointing and migration overhead: heavy deep learning jobs require frequent gradient checkpointing to enable seamless inter-region spatial migration without losing training progress.

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

Arroba, P., Buyya, R., Cárdenas, R., Risco-Martín, J. L., & Moya, J. M. (2024). *Sustainable edge computing: Challenges and future directions*. Software: Practice and Experience, 54(11), 2272–2296. https://doi.org/10.1002/spe.3340

Eilam, T., Joshi, K., & Ramesh, J. (2024). *Reducing datacenter compute carbon footprint by harnessing the power of specialization: Principles, metrics, challenges, and opportunities*. IBM Journal of Research and Development. https://research.ibm.com/publications/reducing-datacenter-compute-carbon-footprint-by-harnessing-the-power-of-specialization-principles-metrics-challenges-and-opportunities

Kim, H., Kim, J., Lee, S., & Choi, Y. (2023). *Carbon-aware computing for sustainable cloud and edge AI systems: Challenges and opportunities*. *arXiv*. https://arxiv.org/abs/2304.00404

Mao, Y., Yu, X., Huang, K., Zhang, Y.-J. A., & Zhang, J. (2023). *Green Edge AI: A contemporary survey*. *arXiv*. https://arxiv.org/abs/2312.00333

Mondal, M. R. H., Islam, M. S., Rahman, M. A., & Hossain, M. S. (2023). *Green cloud computing: Energy-efficient approaches, challenges, and future research directions*. Computers, 12(12), 247. https://doi.org/10.3390/computers12120247