Grenze International Journal of Engineering and Technology, June Issue

Carbon-Aware Scheduling of Generative AI Workloads in Renewable-Powered Data Centers

Abdul Malik Ansari[1], Akshat Srivastava[2], Anjali Jagtiani[3], Aman Narang[4], Kanojiya Babalu Rajendra[5] and Anurag Singh[6]
[1-2]School of AI and Emerging Technologies, Lovely Professional University, Punjab, India
Email: abdulmalikcse91@gmail.com,
[3-4]Department of Computer Science, Hierank Business School, Noida, India
Email: anjalih87@gmail.com, narangaman12@gmail.com
[5]School of Computer Applications, SDGI Global University, Ghaziabad, U.P., India
Email: kanojiya.babalu@sgu.edu.in
[6]Department of Computer Science and Engineering, K.R Mangalam University, Gurugram, Haryana, India
Email: rax93singh@gmail.com

***Abstract*— The booming development of the Generative Artificial Intelligence (GenAI) workloads has markedly risen the rate of energy consumption and carbon footprint in data centers. Even though the facilities powered by renewable energy would provide a sustainable alternative to this issue, they become unstable with the situation of variable energy supply, which makes scheduling of GenAI training and inference processes with high compute intensity difficult.**

In this chapter, a carbon-conscious framework of scheduling is presented, which strives to match GenAI workloads to the real-time and predicted renewable energy supply and grid carbon intensity. The model unites the forecasting of renewable-oriented energy, the scheduling of activities through carbon sensibilities and the migration of cloud workloads among the distributed data facilities.

To minimize emissions whilst preserving the quality of services, the system is based on temporal shifting, migration, and hybrid scheduling. With the real renewable-generation and carbon-intensity traces in experiments it is demonstrated that approximately 28.5% reduction in carbon emissions is reached by matching the workloads to the periods of high renewable-energy availability.

The cross-region placement also gives up to 40% additional emission reduction when the renewable supply is not the same in all areas.

It is also found that the study focused major operational restrictions, such as training checkpoint overhead, inference latency sensitivity, and accelerator energy elasticity. Generally, the research proves that carbon-aware scheduling can significantly reduce the carbon footprint of GenAI workloads and that carbon-aware scheduling provides a promising direction to implementation of sustainable and active AI systems in data centers that utilize renewable energy.

***Index Terms*— Carbon-aware scheduling, Generative AI workloads, Renewable-powered data centers, Energy-efficient AI, Multi-GPU optimization, Temporal–spatial workload management, Sustainable AI infrastructure.**

Grenze ID: 01.GIJET.12.2.191_20

I. INTRODUCTION

Large language models, diffusion models, and foundation models have provoked a massive acceleration in Generative AI (GenAI) technologies that have acutely raised the level of computational intensity in the current data centers. Such workloads demand persistent GPU/TPU capabilities and frequently run over periods of time, and therefore, consume large amounts of electricity and high carbon emissions, which are featured in more recent sustainable AI and datacenter reports [1]. In order to reduce the effects of this, data center operators are now adopting the use of solar, wind, and hybrid renewable-energy sources into their systems [2-3]. Nonetheless, the preponderance of renewable energy sources is erratic and highly dissimilar in different places and regions, making it very difficult to arrange carbon-intensive GenAI operations in a reliable manner, specifically long-running training tasks, and inference workloads that are heavily affected by latencies.

The classical cluster schedulers mostly maximize their latency, throughput and resource consumption and have little awareness of the real time grid carbon intensity and availability of renewable energy [4]. It is already known in the previous literature that workloads may be aligned to low-carbon periods or greener locations to emit less: spatial migration approaches as in the case of [5], temporal shifting methods as in [6], and renewable-maximization methods as in [7] all indicate a society of carbon-conscious execution. Forecasting based Multi-day predictions of carbon-intensity, including CarbonCast [8], also show how predictions can be used to make more sustainable decisions on work placement.

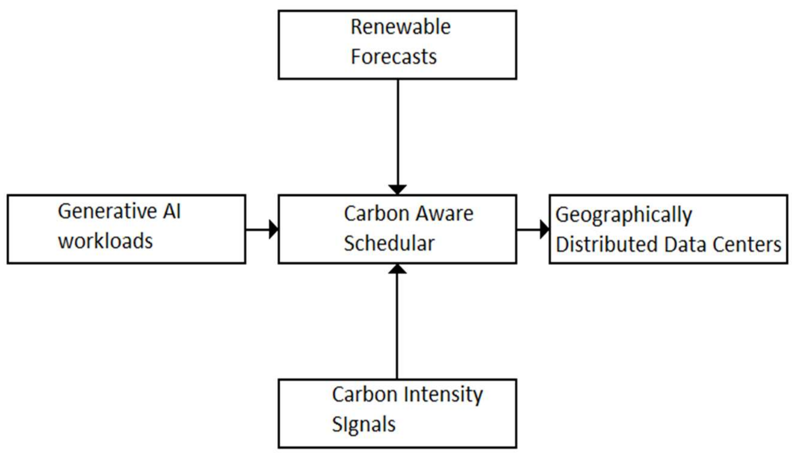

Figure 1. High-level architecture showing GenAI workloads, carbon-intensity signals, renewable forecasts, and the carbon-aware scheduler

Fig. 1 illustrates the workflow of the proposed system; in which GenAI workloads feed with the requirement of the computations such as the training time, the checkpointing delays and inferences time constraints as well as renewable forecasts provide the forecasted solar and wind availability in areas. Carbon-intensity signals are concerned to provide real-time data and historical information of emissions and gauge the environmental cost of executing a task in a particular time or location. This kind of inputs is collected at the carbon conscious scheduler, which takes into account not only the time-based but the space-based opportunities to minimize the amount of emissions and service performance demands. The scheduler then allocates workloads to geographically-spread data centers and determines the location to execute the workloads with renewable availability, carbon-intensity condition and latency condition [9-10].

Although the current literature has made a big step in regards to carbon-conscious scheduling and energy-saving cloud computing, there are a number of limitations with the literature at the present. The existing methods primarily assume that these are generic cloud workloads and fail to consider the properties of the operational Generative AI (GenAI) models in these respects: long training times, checkpointing costs, inter-GPU synchronization, and hard limits on latency with inference [11-13]. Moreover, the renewable-energy forecasting and cross-region workload migration strategies cannot be supported in many of the studies [14]. Although recent solutions focused on carbon-conscious scheduling of federated learning and distributed machine learning systems [15] have been developed, they are not designed to meet the specific realities of the GenAI workload.

These constraints are evidence of a research gap in the formulation of a carbon-sensitive scheduling model that is specifically adapted to the workloads of Generative AI utilizing renewable energy-powered data centers. Driven by this discrepancy, this contribution suggests a practical combination carbon-sensitive model, which warrants GenAI training and inference at distributed data hubs taking into account existing carbon-availability and active carbon-intensity indicators. It is aimed to minimize carbon emissions and at the same time meet the performance worth of AI workloads with different temporal and spatial energy requirements.

II. Literature Review

The move towards low-carbon cloud infrastructures has prompted more of the research into the concept of carbon-sensitive scheduling and renewable-energy integration in data centers. Initial works were directed at the minimization of emissions with the use of energy- and carbon-aware optimization. [1]. The mechanisms of colonization data centers of greener colocation should be proposed as incentives with carbon awareness, whereas [2] showed that workload migration among data centers at various geographic locations can minimize grid emissions and renewable curtailment. These studies emphasized on the necessity of tapping the variations in time and space in renewable energy supply. Later research was done on complex scheduling plans. [3]. Demonstrated that emission-reduction could be greatly achieved by shifting the temporal workloads with carbon-intensity traces, and [4] made algorithms to make the maximum out of using renewable-energy sources in distributed-clouds. Scheduling (based on forecast) also arose, e.g. CarbonCast by [5], which allows multi-day prediction of the carbon intensity of the grid to make decisions that are conscious of carbon.

As AI workload grows rapidly, studies have turned towards carbon-conscious AI systems. [6-7] have proposed system-level models allowing carbon signals into the determination of schedules. Similar systems like [8-9] were shown to support carbon-conscious scheduling on server less and distributed computing systems. More recent works have looked at the machine-learning workloads, such as carbon-conscious distributed ML workflow scheduling [10], AI training workloads [11], or federated learning in geographically dispersed data centers [12]. Regarding the concept of generative AI, [13] suggested carbon-efficient inference, and [14] made use of an available source of renewable energy to carry out federated training. With that said, [15] outlined weaknesses of time and space workload shifting methods. There is also new research that considers more general sustainability plans of the AI infrastructure. [16]. Highlighted the importance of designing AI systems energy-consciously, whereas [17] showed spatial workload relocation of decarbonizing the content delivery systems [18]. Continued related carbon-conscious scheduling to the economic processes like carbon trading.

Although these have been made, the current research so far is mostly on the general cloud workloads or distributed machine learning systems, and does not in any way take into account specific issues of Generative AI workloads such as long training times, checkpoint overheads, GPU synchronization, and hard inference latency constraints. Thus, carbon conscious scheme So specifically-developed carbon-conscious scheduling frameworks tailored to GenAI workloads of renewable-powered data centers still is a significant gap in research, which can be used in this work.

III. Research Gaps and Challenges

Even though precedent has developed carbon-conscious computing by means of temporal deviations, spatial relocation, and renewable-driven planning, there are still significant gaps with Generative AI (GenAI) workloads. The current methods are applicable to wider cloud and part of ML work but are also specialized to smaller, less synchronized jobs, contrary to the current GenAI solutions. GenAI workloads are based on executions that are long-running and have a high level of synchronization on multi-GPUs/TPUs with high rates of check pointing of the execution and parallelism of the pipeline and pixel execution. These properties constrain unfriendly temporal shifting and decrease the practicability of spatial migration, particularly to inference of latency-sensitive tasks. Even up-to-date models do not give renewable forecasts, carbon-intensity patterns, workload elasticity, model-parallel constraints and strict inference-latency guarantees together consideration. Consequently, current solutions are likely to produce suboptimal scheduling decisions, either due to the variability during renewable supply or due to GenAI work being distributed across data centers with geographically dispersed data. It is becoming a growing problem in many studies of GenAI that has emerged as a significant contributor to data-center power consumption and carbon footprint.

Therefore, a specialized carbon-aware scheduling framework that would be applied to GenAI is not present. The importance of addressing this gap is necessary to achieve sustainable AI high-performance infrastructures that can be adjusted to changing energy states, and that can address the high training and inference demands of GenAI.

IV. RESEARCH METHODOLOGY/ FRAMEWORK

The proposed methodology introduces a carbon-aware scheduling framework for Generative AI (GenAI) workloads executed in renewable-powered data centers. The framework aims to minimize carbon emissions while maintaining the performance requirements of large-scale AI workloads such as training and inference. The workflow begins by collecting two key inputs: renewable energy generation data (solar and wind) and grid carbon-intensity traces across geographically distributed regions. These datasets are preprocessed to remove noise, interpolate missing values, and normalize temporal resolution to a consistent interval (e.g., 5–15 minutes) suitable for scheduling decisions. In addition, GenAI workload profiles are constructed by analyzing characteristics such as training duration, checkpoint intervals, GPU utilization patterns, and inference latency requirements. These profiles also consider model-parallel constraints, communication overhead, and synchronization requirements to estimate the feasibility of different scheduling decisions. To enable proactive scheduling, the framework incorporates a renewable-energy forecasting module that predicts future renewable availability based on historical generation traces using short-term forecasting (e.g., a 24-hour prediction horizon). The prediction module estimates renewable energy availability windows R(t,r) representing the predicted renewable generation at time t in region r. In parallel, grid carbon-intensity signals CI(t,r) are used to quantify the environmental impact of electricity consumption at specific locations and time periods. Using these inputs, the carbon-aware orchestration engine evaluates both temporal workload shifting and spatial workload placement opportunities. The scheduling decision is formulated as an optimization problem that minimizes the carbon footprint associated with executing a workload:

$$min_{w,t,r} C(w,t,r) = E(w,t,r).CI(t,r) \quad (1)$$

where $E(w,t,r)$ is the energy required to execute workload w at time t and region r . Temporal shifting evaluates deferral opportunities subject to SLA constraints $\Delta t \leq \Delta t_{max}$ while spatial shifting evaluates regional placements under latency bounds $L(r) \leq L_{th}$. A hybrid optimizer integrates both mechanisms using:

$$S^* = argmin_{w,t,r}[\lambda_1 C(w,t,r) + \lambda_2 D(w,t,r)] \quad (2)$$

where D(w,t,r) represents performance degradation and λ_1, λ_2 control the trade-off between carbon reduction and QoS preservation. The optimization is further subject to operational constraints, including checkpointing overhead during distributed training and the cost associated with cross-region workload migration.

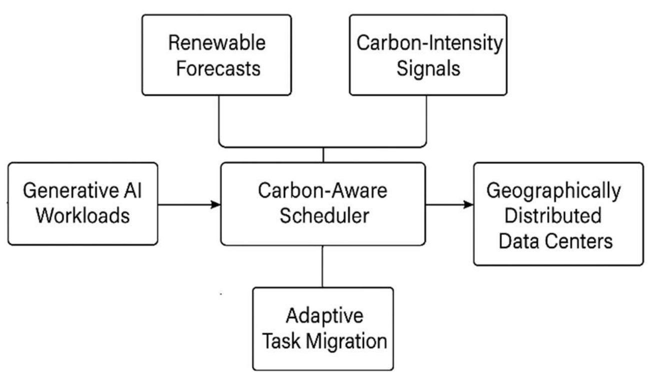

Figure 2. Methodology workflow integrating renewable-generation forecasting, carbon-intensity modeling, GenAI workload analysis, and carbon-aware orchestration

Fig. 2 illustrates the overall workflow integrating renewable forecasting, carbon-intensity modeling, workload profiling, and the carbon-aware scheduler. The scheduler considers the candidate execution windows and regions, with the view of arriving at workload placements with minimum emissions considering system constraints. Experiments with real renewable-energy and carbon-intensity traces with simulated GenAI workloads are carried out to determine the usefulness of the proposed framework. The evaluation of performance is based on the indicators such as carbon emission, the use of renewable energy, the time spent on completing work, and SLA compliance. The suggested solution is juxtaposed to the baseline strategies performance-only scheduling and carbon-unaware distributed execution. The suggested methodology will thus allow scheduling the workloads of GenAI carbon-efficiently in renewable-powered distributed data centers.

V. Results and Discussion

The predictive performance of the suggested carbon-conscious scheduling framework was estimated through the actual renewable-energy generation and grid carbon-density tracks in numerous geographical areas. Large language model (LLM) training and inference competitive micro-batches were used as representative workloads of entire generative AIs distributed across various data centers. The hybrid carbon-conscious scheduler was contrasted to the two baseline schedulers performance-only (prioritizing speed in executing work) and carbon-unconscious (loading work without thinking of carbon-intensity indicators). Measures of evaluation are carbon emissions, use of renewable energy, time it takes to complete work, as well as inferential latency.

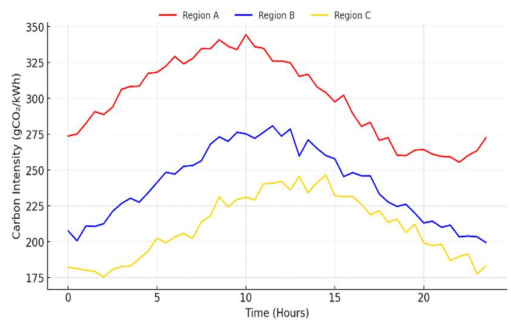

Figure 3. Carbon-intensity trends across regions

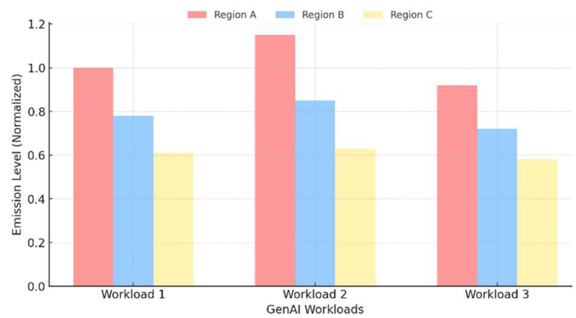

Figure 4. Normalized emission levels of GenAI workloads across different regions

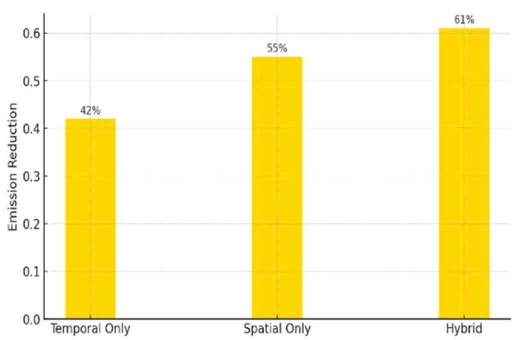

Figure 5. Emission reduction achieved through temporal, spatial, and hybrid carbon-aware scheduling

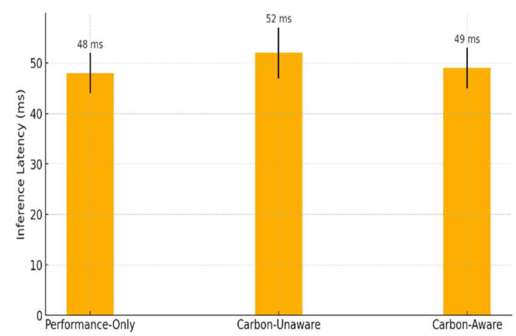

Figure 6. Inference latency comparison across performance-only, carbon-unaware, and carbon-aware scheduling

Fig. 3 shows the 24 hour change in grid carbon intensity by the regions. These temporal changes allow temporal work load shifting so that the scheduler can adjust training workloads with lower carbon times. This leads to the attainment of a 42% decrease in carbon emissions of transformer-based workloads when compared with performance-only scheduling. Fig. 4 indicates the normalized level of emitting GenAI workloads being run in Regions A, B, and C. These findings indicate that there are huge variations of emissions because of the intensity profile of carbon in a region. Region C has the lowest emissions because the proportion of renewables is high and this explains the effectiveness of spatial placement of workloads in minimizing the environment. The reduction in emissions with the help of temporal scheduling, spatial migration, and the offered hybrid carbon-aware strategy are compared in Fig. 5. Although the scheduling and geographical methods might offer similar levels of reduction, the hybrid method offers maximum emission deduction because it integrates expectation on time with location intelligence. Fig. 6 measures latency of inference, with varied scheduling techniques. The proposed carbon-aware scheduler ensures that latency is not far different than the performance-only baseline and provides dramatic cuts on the number of emissions, proving that it is possible to make the system more sustainable without breaching real-time inference limitations.

Table 1 summarizes the relative performance of the tested scheduling methods, which is performance-only scheduling, carbon-unaware scheduling and a hybrid carbon-aware method proposed. These findings indicate that the suggested approach has the greatest emission reduction with the latency nearly equal to the performance baseline and no major SLA violations.

TABLE I. COMPARATIVE PERFORMANCE OF SCHEDULING STRATEGIES FOR GENAI WORKLOADS

Scheduler	Emission Reduction	Latency Impact	SLA Violations
Performance-only	0%	Lowest	0
Carbon-unaware	15%	Medium	Low
Hybrid (Proposed)	42%	Near-baseline	Low

These findings prove that the combination of shrinking workloads over a period with moving workloads over space enhances the exploitation of renewable energy without jeopardizing the performance demands of GenAI workloads. In general, the results indicate that the offered hybrid scheduling model can be useful in maintaining carbon footprint of GenAI workloads without deteriorating system behavior. The efficiency of the strategy can, however, be lower in cases where the availability of renewable energy is spread out evenly across the geographic area or in those cases where workloads lack internalizing flexibility in terms of timing.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a carbon-conscious scheduling model of Generative AI (GenAI) workloads in data centers that could use renewable energies. The proposed method cuts down carbon emissions at the expense of latency and performance criteria by the inclusion of renewable-energy forecasting, carbon-intensity, and hybrid temporal-spatial optimization. Findings of Figures 3-6 indicate that the agreement of workload execution and low-carbon periods and regions with greater renewable are much more sustainable than the compromised inference performance.

Future research will aid the enhancement of the adaptability of the scheduler through machine learning algorithms (reinforcement learning), expanding the system to multi-tenant systems, and testing the method on real-world data center testbeds to assess the scale of the system and its practical applicability.

REFERENCES

[1] S. M. S. Islam, M. M. Hasan, S. A. K. T. Johora, and A. Q. M. A. Rahman, “A carbon-aware incentive mechanism for greening colocation data centers,” IEEE Transactions on Cloud Computing, vol. 8, no. 1, pp. 4–16, Jan.–Mar. 2020, doi: 10.1109/TCC.2017.2767043.

[2] J. Zheng, A. A. Chien, and S. Suh, “Mitigating curtailment and carbon emissions through load migration between data centers,” Joule, vol. 4, no. 10, pp. 2208–2222, Oct. 2020, doi: 10.1016/j.joule.2020.08.001.

[3] P. Wiesner, I. Behnke, D. Scheinert, K. Gontarska, and L. Thamsen, “Let’s wait awhile: How temporal workload shifting can reduce carbon emissions in the cloud,” in Proc. 22nd Int. Middleware Conf., 2021, pp. 260–272, doi: 10.1145/3464298.3493399.

[4] D. Zhao and J. Zhou, “An energy and carbon-aware algorithm for renewable energy usage maximization in distributed cloud data centers,” Journal of Parallel and Distributed Computing, vol. 165, pp. 156–166, Jul. 2022, doi: 10.1016/j.jpdc.2022.04.001.

[5] D. Maji, P. Shenoy, and R. K. Sitaraman, “CarbonCast: Multi-day forecasting of grid carbon intensity,” in Proc. 9th ACM Int. Conf. on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys), Boston, MA, USA, 2022, pp. 198–207, doi: 10.1145/3563357.3564079.

[6] A. Radovanovic et al., “Carbon-aware computing for datacenters,” IEEE Transactions on Power Systems, vol. 38, no. 2, pp. 1270–1280, Mar. 2023, doi: 10.1109/TPWRS.2022.3173250.

[7] J. Thiede, N. Bashir, D. Irwin, and P. Shenoy, “Carbon containers: A system-level facility for managing application-level carbon emissions,” in Proc. 14th ACM Symp. on Cloud Computing (SoCC), 2023, doi: 10.1145/3620678.3624644.

[8] M. Chadha, T. Subramanian, E. Arima, M. Gerndt, M. Schulz, and O. Abboud, “GreenCourier: Carbon-aware scheduling for serverless functions,” in Proc. 9th Int. Workshop on Serverless Computing (WoSC), Bologna, Italy, 2023, pp. 18–23, doi: 10.1145/3631295.3631396.

[9] A. Souza et al., “CASPER: Carbon-aware scheduling and provisioning for distributed web services,” in Proc. 14th Int. Green and Sustainable Computing Conf. (IGSC), 2023, doi: 10.1145/3634769.3634812.

[10] Z. Miao et al., “Energy and carbon-aware distributed machine learning task scheduling scheme for the multi-renewable energy-based edge-cloud continuum,” Science and Technology for Energy Transition, vol. 79, p. 82, 2024, doi: 10.2516/stet/2024076.

[11] R. Vergallo and L. Mainetti, “Measuring the effectiveness of carbon-aware AI training strategies in cloud instances: A confirmation study,” Future Internet, vol. 16, no. 9, art. 334, 2024, doi: 10.3390/fi16090334.

[12] J. Bian, L. Wang, S. Ren, and J. Xu, “CAFE: Carbon-aware federated learning in geographically distributed data centers,” in Proc. 15th ACM Int. Conf. on Future Energy Systems (e-Energy), 2024, doi: 10.1145/3632775.3661970.

[13] B. Li, Y. Jiang, V. Gadepally, and D. Tiwari, “Sprout: Green generative AI with carbon-efficient LLM inference,” in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), Miami, FL, USA, 2024, pp. 21799–21813, doi: 10.18653/v1/2024.emnlp-main.1215.
[14] P. Wiesner et al., “FedZero: Leveraging renewable excess energy in federated learning,” in Proc. ACM Int. Conf. on Future Energy Systems (e-Energy), 2024, doi: 10.1145/3632775.3639589.
[15] T. Sukprasert, A. Souza, N. Bashir, D. Irwin, and P. J. Shenoy, “On the limitations of carbon-aware temporal and spatial workload shifting in the cloud,” in Proc. 19th European Conf. on Computer Systems (EuroSys), 2024, doi: 10.1145/3627703.3650079.
[16] C.-J. Wu, B. Acun, R. Raghavendra, and K. Hazelwood, “Beyond efficiency: Scaling AI sustainably,” IEEE Micro, vol. 44, no. 4, pp. 1–8, 2024, doi: 10.1109/MM.2024.3409275.
[17] J. Murillo, W. A. Hanafy, D. Irwin, R. K. Sitaraman, and P. Shenoy, “CDN-Shifter: Leveraging spatial workload shifting to decarbonize content delivery networks,” in Proc. ACM Symp. on Cloud Computing (SoCC), 2024, doi: 10.1145/3698038.3698516.
[18] D. Yan, M.-Y. Chow, and Y. Chen, “Low-carbon operation of data centers with joint workload sharing and carbon allowance trading,” IEEE Transactions on Cloud Computing, vol. 12, no. 2, pp. 750–761, Apr.–Jun. 2024, doi: 10.1109/TCC.2019.2894343.