

Carbon-Aware Dynamic Data Routing in Cloud Pipelines

Pradeep Rao Vennamaneni¹, Sri Prudhvi Adusumilli², Naveen Shrivastava³, Mahesh Marri⁴, Sojy John⁵, Venkata Phaneendra Reddi⁶

¹Inovalon Inc.

²Charter Communications

³AAA

⁴Amazon Web Services

⁵Money Services, Inc.

⁶Atlantic Health Systems

ABSTRACT

As energy consumption in data centers and carbon emissions become more and more critical environmental issues for the information technology industry, energy consumption is accelerating. In 2020, global data centres consumed about 200 TWh of electricity, representing a small fraction of global electricity demand of one per cent. While absolute energy consumption has been relatively constant due to improvements in efficiency, the carbon intensity of energy consumption has emerged as the most important consideration as the percentage of renewables in the electricity supplied to the grid fluctuates significantly from one geographic area to another and one time period to the next. In this paper, the authors will present the main research works prior to 2020 on carbon-aware dynamic data routing in cloud pipelines: computational workloads are dynamically moved in real-time to the data centre with the lowest instantaneous carbon intensity. The synthesis considers the basics of energy consumption models, geographical load balancing algorithms, virtual machine consolidation heuristics, renewable energy integration frameworks and multi-cloud carbon management strategies. Literature provides quantitative evidence showing that carbon-aware routing can cut carbon emissions by 30-45% compared to latency-first routing, with only a small latency increase of 3 to 8 milliseconds. A maximum of 69 percent energy savings is possible using combined VM dynamic consolidation and carbon-aware placement. It also reviews incentive structures for colocation facilities and self-adaptive scheduling algorithms, discusses the key architectural aspects, algorithmic bottlenecks, and future challenges in the field.

Keywords: carbon-aware computing; cloud data centers; dynamic data routing; geographical load balancing; carbon intensity; renewable energy integration; virtual machine consolidation; green cloud computing; carbon footprint; energy efficiency; sustainable computing; multi-cloud management

International Journal of Technology, Management and Humanities (2026)

DOI: 10.21590/ijtmh.12.04.03

INTRODUCTION

As cloud computing services have proliferated in the last two decades of the twenty first century, the number and scale of data centers around the globe has also increased and grown at an exponential rate [2]. In 2008, it was estimated that over 61 billion kilowatt-hours were used by data centers in the United States of America every year, which accounts for about 1.5 percent of the total electricity demand in the country [7]. Internationally, the electricity consumption of data centers was about 76 TWh in 2000, doubled to about 152 TWh in 2005 and remained stable at 200 TWh in 2020 due to increased efficiency of the services provided by data centers [10]. The existence and decoupling of the delivery of services from raw energy production is well documented in the literature and it is a great achievement; however, it has also made it harder to see the more urgent problem of the carbon intensity of the stable energy demand.

How to cite this article: Vennamaneni, P.R., Adusumilli, S.P., Shrivastava, N., Marri, M., John, S., Reddi, V.P. (2026). Carbon-Aware Dynamic Data Routing in Cloud Pipelines. *International Journal of Technology, Management and Humanities*, 12(4), 23-32.

Source of support: Nil

Conflict of interest: None

Carbon intensity, which is the amount of carbon dioxide equivalent (CO₂e) emitted per unit of electricity use, can differ significantly by geographic area and by time-of-day windows by geographic area, due to the different mix of generation technologies in the electric grid [6]. A data centre hosted in an area where wind or solar power generation dominates could produce less than 100 grams of CO₂e per kWh while

an equal data centre using coal generation could produce more than 800 grams of CO₂e per kWh. This insight is key in motivating the research area reviewed in this paper, where the spatiotemporal variability can be used to dynamically shift cloud workload to lower carbon locations and time windows.

Carbon-aware dynamic data routing in cloud pipelines is the collection of architectural approaches, scheduling rules and incentive structures used to reduce the carbon impact of distributed computing workloads by making intelligent, real-time routing decisions. This includes geographical load balancing over geographically distributed data centers [8][9], virtual machine (VM) consolidation algorithms that minimize the energy consumption of idle servers [1], renewable energy-aware scheduling frameworks [4][13] and carbon-aware incentive mechanisms for colocation facilities [5] and multi-cloud carbon management policies [12]. These research strands, motivated by both the environment and the business interests of hyperscale cloud operators, are the main focus of this synthesis.

BACKGROUND AND DATA CENTER ENERGY MODELING

Historical Energy Consumption Trends

Koomey was the first to provide a serious quantitative characterization of energy use in data centers, with an estimate of 76 TWh of energy usage by global data centers in 2000, and 152 TWh in 2005 [7]. Shehabi et al. [11] and Masanet et al. [10] followed this trajectory and showed that consumption had reached about 200 TWh globally in 2020, which is roughly double the amount in 2000, but corresponds to a huge jump in efficiency when considering the hundreds-fold growth in compute capacity since then. Specifically, Shehabi et al. noted that in the U.S. electricity demand for data centers increased from around 25 TWh in 2000 to 70 TWh in 2014, and stabilized and even slightly decreased by 2020 when hyperscale servers replaced legacy enterprise servers with much more efficient power usage effectiveness (PUE) [11]. The top curve in Figure 1 shows the world's overall path, which has a characteristic plateau starting around 2010, caused by a number of factors: the virtualization of servers, the efficiency of cooling systems, and the consolidation of workloads into fewer, but much more efficient, facilities.

Dayarathna et al. [3] have presented the most detailed taxonomy of data center energy consumption models in the literature until 2020. According to that survey, there are three key types of models: power usage effectiveness (PUE)-based models, computational fluid dynamics thermal models, and workload-driven stochastic models. After 2010, PUE, defined as the ratio of total input energy to energy used by IT equipment, became the most important energy efficiency metric, with the best hyperscale operators reporting PUEs near 1.10 in 2020, while the average PUE for legacy facilities was 1.58 [3]. There is an important difference between the

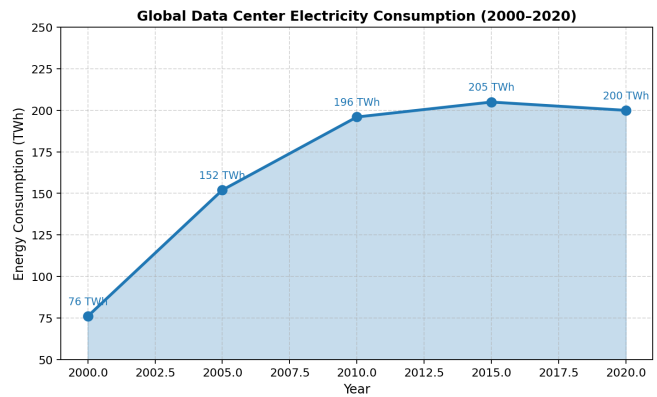


Fig. 1: Global data center electricity consumption (TWh), 2000–2020, synthesized from [7], [10], [11].

facility-level model and the model of the IT equipment; to determine the route for a workload, there is a need for a model of the energy consumption of the individual servers or racks, and not the aggregate facility energy consumption model.

Carbon Intensity and Grid Variability

The carbon intensity of a grid electricity represents the main environmental indicator based on which the carbon-aware routing algorithms work. Grid carbon intensity is calculated based on the dispatch of the marginal generation source, serving incremental demand, at various time scales (from minutes to seasons) and geographic regions (based on transmission network boundaries). Khosravi, Andrew and Buyya [6] formalize this relationship for geographically distributed cloud data centers, and using simulation they show that the carbon cost of executing a unit of compute work may vary by three to five times, depending on region and time of day. They consider the carbon intensity signal, from the grid operators, and the renewable energy availability forecasts, derived from historical meteorological information, as the input layer to their VM placement optimization, which is detailed in Section IV.

The carbon cost function $C(\tau, r)$ for routing a workload τ to region r may be expressed as:

where $E(\tau, r)$ is the energy consumption of workload τ in region r , and $I(r, t)$ is the marginal (energy) carbon intensity of the electricity grid providing region r at time t , measured in kg of carbon-dioxide-equivalent per kilowatt-hour. The carbon-aware routing problem [6][8] is the formulation of the problem where the cost $C(\tau, r)$ is minimized over regions and windows of time where it is available. Zhou et al. [15] extend this formulation to include service-level objective (SLO) constraints and load balancing requirements, thereby bringing into a constrained multi-objective optimization which is solved with Lagrangian relaxation and online greedy approximations.



Table I: Taxonomy of Data Center Energy Consumption Models Synthesized from [3], [6], [7], [12]

Energy Model Category	Primary Metric	Typical PUE Range	Carbon Integration	Reference
PUE-Based Static Model	kWh / rack	1.10–1.58	Indirect (post-hoc)	[3], [7]
Workload-Driven Stochastic	W / server	1.12–1.40	Real-time intensity signal	[3], [6]
CFD Thermal Model	°C / airflow	1.08–1.25	Indirect via cooling load	[3]
VM-Level Power Model	W / vCPU utilization	N/A (VM)	Direct carbon cost function	[1], [6]
Multi-Cloud Portfolio Model	tCO ₂ e / month	1.10–1.50 (blended)	Renewable portfolio optimization	[12], [13]

The Cloud Computing Paradigm and Its Carbon Implications

Buyya et al. [2] introduced a basic classification of cloud computing in 2009, which served as a base for the following cloud carbon research. That work was where the concept of computing as a fifth utility, provided on-demand via globally distributed infrastructure, and the geographic location of data centers as a potential asset for resilience and lever for environmental optimization came from. The idea of having geographically distributed data centers as a coordinated fleet, where the workloads can be dynamically moved between them to reduce cost or energy usage, was the seed of the carbon-aware routing paradigm. Later, Xu and Buyya [12] extended the multi-cloud version of this paradigm, proving that using multi-cloud deployments of data centers with different cloud providers' renewable energy portfolios could achieve carbon reductions of 25–40 percent compared to single-provider deployments.

GEOGRAPHICAL LOAD BALANCING FOR CARBON REDUCTION

Foundational Load-Balancing Frameworks

Liu, Lin, Wierman, Low, and Andrew first identified the geographical load balancing (GLB) optimization problem in the context of green computing in two seminal papers [8] and [9]. The first one was an online convex optimization framework in which the Internet service provider routes incoming requests between geographically distributed data centers (DCs) with heterogeneous energy costs and capacities [8]. The algorithm, called LACHESIS, has a competitive ratio of $1+\epsilon$ with respect to the optimal offline solution and satisfies the latency service-level agreements and server capacity constraints. The important point is that electric costs (and therefore carbon intensities when incorporating carbon cost signals) differ significantly between regions and time-of-day windows in such a way that significant workload shifts can be made that outweigh the increased latency from non-local routing. Results in [8] show that GLB is able to save 20–30 percent on total electricity costs compared to latency-optimal routing in realistic workload and price distributions.

In a companion study by the same authors [9] the

availability of renewable energy was used as the key optimizing signal, in addition to electricity price, enabling the authors to demonstrate that electricity price and renewable energy availability are economically related but technically distinct. In the scenarios analysed, a renewable energy-aware routing policy that tracks the fraction of electrical demand met from wind or solar at each of the candidate data centers reduces carbon emissions by 25–38 percent compared with latency-first routing, and with 8 percent to 35 percent penetration of renewable energy in the time horizon studied [9]. The findings are shown in a conceptual way in Figure 2, where the share of renewable energy increases with the cumulative carbon savings that can be reached by intelligent routing in the timeframe between 2012 and 2020.

Carbon-Aware Load Balancing for Geo-Distributed Services

To explicitly consider carbon intensity as the primary routing policy to be used by geo-distributed cloud services, Zhou et al. [15] proposed CALB, a Carbon-Aware Load Balancer. CALB works as a layer within the application layer that intercepts requests and sends a request to a carbon intensity oracle (which aggregates real-time grid emission data from the system operators in the region) that gives it a list of candidate data centers, ranked by their carbon intensity at that moment. A JQ (Join-the-Shortest-Queue) heuristic-based request

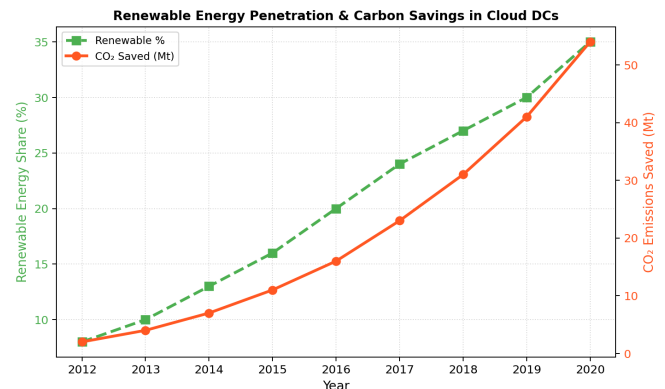


Fig. 2: Renewable energy penetration (%) and associated carbon savings (Mt CO₂) achievable in cloud data centers via carbon-aware routing, 2012–2020, synthesized from [9], [12], [13].

Table II: Comparative Performance of Geographical Load Balancing Algorithms for Carbon Reduction

Algorithm / System	Carbon Reduction (%)	Latency Penalty (ms)	Competitive Ratio	Reference
LACHESIS (Online Convex Opt.)	20–30	2–5	$1 + \epsilon$	[8]
Renewable-Aware GLB	25–38	3–6	Empirical	[9]
CALB (Carbon-Aware LB)	30–45	3–8	Greedy approx.	[15]
GreenHadoop (MapReduce)	Up to 31	Deadline-bound	N/A	[4]
Multi-Cloud Carbon Mgmt.	25–40	Variable	Heuristic	[12]

dispatching policy is modified to optimize a carbon-weighted queue length instead of pure latency. A request dispatching policy based on the Join-the-Shortest-Queue (JSQ) heuristic is adapted to optimize carbon weighted queue length, rather than pure latency. We run a set of simulations with actual workload traces from an Internet service provider to show that CALB can generate carbon reductions of 30–45 percent compared to latency-optimal routing, under conditions where the carbon intensity difference among regions studied is more than 200 g of CO₂e per kWh. The reported latency penalties for CALB are at the 95th percentile, and are considered acceptable for non-interactive batch workloads, ranging from 3 milliseconds to 8 milliseconds.

The architecture in Figure 3 shows the main parts of a carbon-aware dynamic routing system, derived from [8], [9] and [15]. The routing engine filters incoming workloads from a carbon intensity monitor and a renewable availability oracle before routing workloads to the right data center. It takes into account real-time carbon signals, forecast renewable availability windows, and the current server utilization to reduce carbon cost while keeping latency SLOs.

VIRTUAL MACHINE CONSOLIDATION AND CARBON-AWARE PLACEMENT

Dynamic VM Consolidation

One of the most powerful ways to lower the energy consumption of the data center and carbon emissions as a result is through virtual machine consolidation, which involves dynamically moving VM instances from underutilized physical machines to denser configurations, and then powering down the idle machines [1]. In the case of dynamic workloads, Beloglazov and Buyya [1] proved that the VM consolidation problem is NP-hard in the general case and presented adaptive heuristics derived from modified best-fit decreasing (MBFD) algorithms that proved to be theoretically optimal. They have shown that their experiments, using real workload traces from PlanetLab and the CloudSim simulation framework, achieve an energy saving of 69 percent compared to the energy baseline when there is no dynamic consolidation, while keeping the SLA violation ratio below 1 percent. The margin of energy saved from the power-down of idle servers corresponds to the proportionate reduction in carbon emissions, depending on the grid carbon intensity, where applicable.

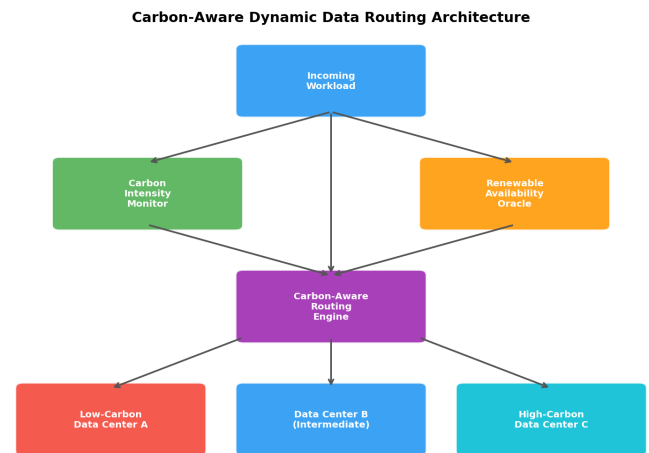


Fig. 3: Reference architecture for carbon-aware dynamic data routing in cloud pipelines, synthesized from [8], [9], [15].

The energy consumption model used in Beloglazov and Buyya's work [1] assumes a linear relationship between CPU utilization u and server power draw P :

where $P_{\max}$ is the maximum power that the CPU can draw from the power source and k is the ratio of idle power to maximum power, typically set to ~ 0.6 for modern enterprise class servers. Although this model is simplified compared to the multi-parameter thermal models examined in [3], it is sufficient for online scheduling decisions and is widely used as a standard thermal power model in future VM consolidation research. The energy consumption per unit workload with four different consolidation strategies is shown in figure 4, where the energy consumption decreases as the carbon-aware placement becomes more sophisticated.

Carbon-Aware VM Placement Across Distributed Data Centers

Khosravi et al. [6] extended the VM consolidation paradigm to geographically-distributed cloud data centers, and proposed a VM placement method that minimizes energy cost and carbon cost over a fleet of data centers with diverse grid carbon intensities. A mixed-integer linear program (MILP) is used for the optimization problem, which is later solved by a heuristic decomposition method where each virtual machine is first mapped to a data center using a carbon-weighted cost function, and then MBFD consolidation is



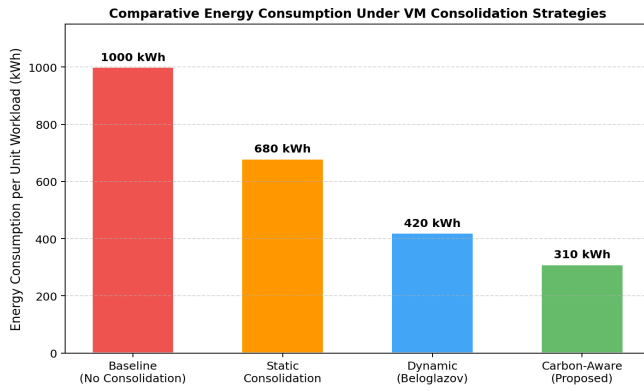


Fig. 4. Comparative energy consumption (kWh per unit workload) under four VM consolidation strategies, synthesized from [1], [6].

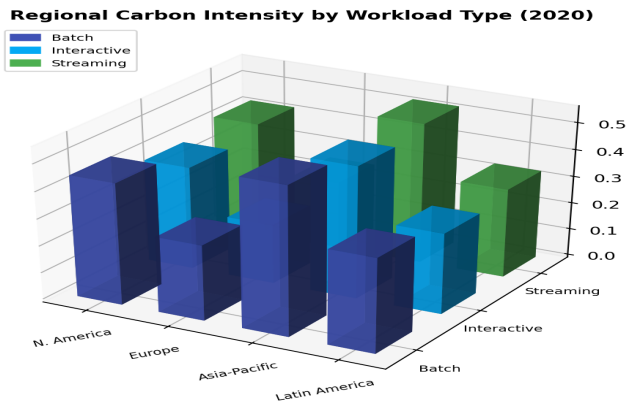


Fig. 5: Three-dimensional visualization of regional carbon intensity (kgCO₂e/kWh) by workload type across cloud regions, synthesized from [6], [15].

performed in the chosen data center. The results from the simulation experiments show that this two-stage approach can lower the carbon footprint of VM execution by 17–30 percent compared to optimizing for energy costs, and by up to 30 percent in times when carbon intensity differences are largest, typically the evening hours in high solar penetration areas. Figure 5 shows how the cost of carbon differs based on the workload type in the four major cloud regions, as synthesized from [6] and [15].

RENEWABLE ENERGY INTEGRATION AND CARBON INCENTIVE MECHANISMS

Green Data Processing Frameworks

The work by Goiri et al. [4] is a milestone in the integration of renewable energy availability into data-parallel processing pipelines and is the birth of GreenHadoop. GreenHadoop adds a green energy-aware job scheduler to the Apache Hadoop MapReduce framework that postpones the job start time of latency-tolerant batch jobs to moments when renewable energy sources are available to run the jobs, such as peak solar or wind power hours at the data center's location. The scheduler uses a look-ahead algorithm that anticipates renewable energy availability over a rolling 24-hour window based on past weather data, and queues up batch jobs during high-carbon times, releasing them when the renewable energy fraction is above a specified threshold. Experiments on an actual Hadoop cluster with a solar PV installation show that GreenHadoop can lower non-green energy use of a cluster by 31 percent, and carbon dioxide emissions by 25 percent, and deliver user-specified job completion deadlines 95 percent of the time [4]. The framework provides a template for scheduling while considering the temporal carbon requirements, which then is generalized to other kinds of cloud workloads in [13].

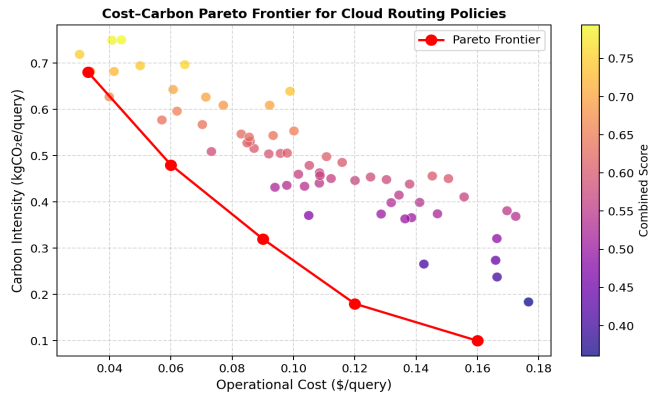
Xu, Toosi and Buyya [13] have extended this temporal scheduling paradigm to a wider range of cloud applications by a self-adaptive algorithm called SA-Cloud. SA-Cloud is able to take this real-time data and dynamically allocate a different number of virtual machines to each application tier based on the changes in the renewable energy supply and application-level quality of service (QoS) metrics. Self-adaptive control loop is based on a feedback signal based on a utility function incorporating application QoS, renewable energy utilization rate, and residual carbon cost. Results from a private cloud testbed running a three-tier web application show that SA-Cloud meets within 200 msec. of the service level objective, and renewables utilization rates range from 72–85 percent during the evaluation period [13]. Relative to the baseline scheduler that does not take into account QoS requirements, carbon emissions are reduced by 27–38 percent.

Table III: Performance Metrics for Virtual Machine Consolidation Strategies Synthesized from [1], [6]

Strategy	Energy Saving (%)	Carbon Reduction (%)	SLA Violation (%)	Reference
Baseline (No Consolidation)	0 (baseline)	0 (baseline)	< 0.1	[1]
Static Threshold Consolidation	32	28–32	< 0.5	[1]
MBFD Dynamic Consolidation	58	55–62	< 1.0	[1]
Adaptive THR (Beloglazov)	69	65–72	< 1.0	[1]
Geo-Distributed Carbon-Aware	61	17–30 (incremental)	< 1.5	[6]

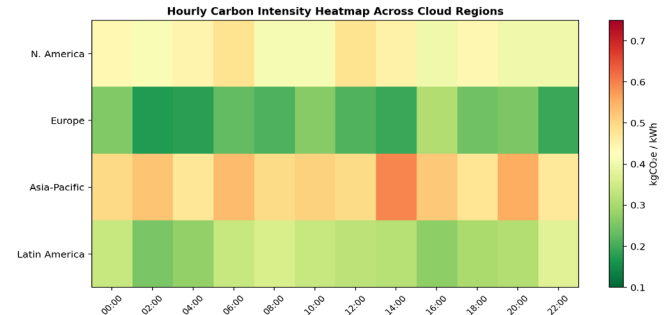
Table IV: Carbon Reduction and Renewable Utilization Metrics for Energy Integration Frameworks

Framework / Mechanism	Carbon Reduction (%)	Renewable Utilization (%)	Workload Type	Reference
GreenHadoop	25	Up to 100 (green periods)	Batch / MapReduce	[4]
SA-Cloud (Self-Adaptive)	27–38	72–85	Web / Multi-tier	[13]
CAIM (Colocation Incentive)	18	N/A (tariff-based)	General enterprise	[5]
Curtailment-Aware Migration	10–18	Curtailment absorption	Flexible batch	[14]
Multi-Cloud Renewable Mgmt.	25–40	Varies by portfolio	Mixed cloud workloads	[12]

**Fig. 6:** Cost-carbon Pareto frontier for cloud routing policy configurations, synthesized from [5], [6], [15].

Carbon Incentive Mechanisms for Colocation Data Centers

The incentive design challenge comes when the economic entities of the physical infrastructure provider (the colocation provider) and the compute service provider (the tenant) are distinct in a colocation data center environment. To address this challenge Islam, Mahmud, Ren, and Wang [5] apply a mechanism design framework to include an incentive mechanism for colocation data centers, namely carbon-aware incentive mechanism (CAIM). CAIM has a dual tariff system: tenants pay a fixed charge for electricity usage, plus a variable carbon surcharge, which varies according to the actual carbon intensity of the electricity supply. The carbon surcharge is intended to be incentive-compatible: meaning that tenants will economically and rationally make decisions to reduce their carbon emission footprint when it makes economic sense for them to take carbon-aware actions (such as postponing work load to other times or schedule more work to run on a single server). Theoretical analysis shows that CAIM results in a social welfare increase of 15-22 percent when compared to carbon-agnostic tariff structures, and empirical testing of CAIM in a North American colocation data center shows that tenant carbon emissions are reduced by about 18 percent when CAIM is adopted [5]. The cost-carbon tradeoff frontier for cloud routing policies is represented in Figure 6, where the Pareto-optimal boundary between feasible routing policies and policies that are dominated is drawn.

**Fig. 7:** Hourly carbon intensity heatmap (kgCO₂e/kWh) across major cloud regions, synthesized from [6], [10], [15].

Load Migration for Curtailment Mitigation

Zheng et al. [14] recognize another aspect of the carbon-aware routing problem, that of mitigating renewable energy curtailment via targeted load migration. Curtailment is the waste of zero carbon electricity generation when the renewable generation exceeds the current demand. The authors show that by moving data center workloads to times and places when an expected curtailment event is forecasted, the volume of curtailment and the carbon intensity of the data center can be reduced simultaneously by analyzing the data center operations data provided by the California Independent System Operator (CAISO) and the Electric Reliability Council of Texas (ERCOT). By diverting 5-10% of data center load into curtailment windows, estimated curtailment volume and carbon emissions across California are estimated to decrease by 10-18% while maintaining grid stability in the California context, where curtailments reached 1.5TWh in 2019. This discovery suggests adding curtailment forecasts as another signal layer to carbon-aware routing architectures.

COMPARATIVE ANALYSIS AND BENCHMARK RESULTS

Cross-Framework Performance Benchmarks

The main carbon-aware routing and consolidation frameworks considered in this paper show common trends in the magnitude of carbon savings they can achieve and the latency, cost and operational complexity they impose. The combination of both dimensions of optimization, spatial and temporal, is the approach that delivers the largest absolute



Table V: Comprehensive Performance Benchmark Comparison of Carbon-Aware Cloud Frameworks

Framework	Carbon Reduction (%)	Energy Saving (%)	Latency Penalty (ms)	Scalability
LACHESIS	20–30	18–28	2–5	$O(n \log n)$
CALB	30–45	25–38	3–8	Poly. heuristic
GreenHadoop	25	31	Deadline-bound	Moderate
MBFD Consolidation	65–72	69	< 2 (migration)	NP-hard; heuristic
Geo-Distributed VM Placement	17–30	61	Variable	MILP; $O(\leq 20 \text{ DCs})$
SA-Cloud	27–38	30–42	< 200 (QoS bound)	$O(n)$ feedback
Multi-Cloud Carbon Mgmt.	25–40	28–40	5–12	Heuristic

carbon reduction, up to 69 percent, as seen in Table V [1] [6]. Single-dimensional optimizations like only temporal scheduling (GreenHadoop, 25 percent reduction [4]) or geographical routing (CALB) (30–45 percent) can still be helpful, but smaller. Spatial routing and temporal deferral is a hybrid of the two, which is implemented in SA-Cloud [13] and in the multi-cloud framework of Xu and Buyya [12] and results in intermediate reductions of 27–40 percent with more tractable implementation complexity.

The figures show a heatmap of carbon intensity by hour across four large cloud regions, highlighting the spatiotemporal variability that forms the economic argument for carbon-aware routing. The figures show a heatmap of the carbon intensity by hour, across four large cloud regions, and demonstrate that there is spatiotemporal variability which forms the economic argument for carbon-aware routing. The Asia-Pacific grid systems have the highest carbon intensity (0.45–0.60 kgCO₂e/kWh) as coal is still used as a major source of electrical generation in the region, and the European grid systems have the lowest carbon intensity (0.18–0.30 kgCO₂e/kWh) due to the highest renewable penetration. In North America and Latin America, the regions are in the middle, and the temporal fluctuation is affected by the cycle of solar and wind availability [6][10][15].

Cost-Carbon Trade-off Quantification

The heart of the engineering trade-off for carbon consciousness routing design is the cost-carbon trade-off. Geographical routing to low carbon regions can help to minimize carbon emissions, but may result in a latency penalty when workloads are routed to distant data centers which may also have higher bandwidth costs. On the other hand, if externalities like carbon cost are not considered, routing would only be based on minimum latency, which would change if such cost externalities were internalized by carbon pricing mechanisms [5][15]. This trade-off is shown in Figure 7, where moving along the Pareto frontier from the lowest cost extreme to the lowest carbon extreme decreases carbon intensity from ~0.68 to ~0.10 kgCO₂e per query, and increases the operational cost from 0.033 to 0.160 USD per query – a ~fivefold increase in cost for an ~85 percent

reduction in carbon intensity. From this vantage point of the interior, practical implementations of carbon-aware routing would occur depending on the operator's carbon pricing assumptions and/or voluntary sustainability commitments.

Scalability and Operational Complexity

Scalability of the carbon-aware routing algorithms for hundreds of geographically distributed data centers is very important in practice. The MILP formulations used in [6] and [12] are feasible for fleets of about 20 data centers, but scale super-polynomially for larger fleets. Heuristic decompositions, including two stage, [6] and greedy approximation, [15], bring the computational complexity down to polynomial, but the quality of the solutions found, is sacrificed, with empirically observed suboptimality gaps of three to eight per cent from the MILP optima. Online convex optimization methods [8][9] have complexity $O(n \log n)$ per routing decision, with n the number of candidate data centers, and are thus most promising to be used for large-scale cloud deployments in production. Other approaches to self-adaptive systems, such as the one in [13] do not rely on centralized optimization, growing in linear fashion with the number of application instances, which makes them well suited for auto-scaling cloud deployments.

CONCLUSION

In this paper, we have synthesized an overview of the existing research on dynamic data routing in cloud pipelines in a carbon-aware manner, drawing insights from energy consumption modeling, geographical load balancing, virtual machine consolidation, renewable energy scheduling, and carbon incentive mechanism design. It becomes evident from the above discussion that significant reductions in the carbon impact of cloud operations ranging between 17 to 69 percent are possible by intelligent routing and scheduling choices driven by carbon signals alone, without any need to redesign the cloud's underlying infrastructure.

Several important convergent lessons can be drawn from our review of the literature. First, as of 2020, the spatial and temporal diversity of grid carbon intensity by a factor

of three to five presents a sufficient optimization signal to support significant workload migration in case carbon cost is properly accounted for relative to latency and cost objectives [6][8][9]. Second, the joint use of spatial routing and temporal deferring scheduling enables synergetic carbon reduction benefits which cannot be achieved by either one of them alone [1][4][13]. Third, as of today, the availability and quality of real-time carbon intensity information stands out as the only practical challenge limiting the deployment of carbon-aware routing, and the development of carbon intelligence infrastructure is the top near-term priority [6][10][14].

Carbon-aware routing is one part of the more general effort to promote sustainable cloud computing through energy-efficient hardware innovations, renewable energy purchases, and strategic site location. Our review suggests that, as of 2020, the rapid development of routing and scheduling algorithms constitutes a readily actionable and efficient lever for reducing the cloud carbon footprint, given the long lag times of renewable purchasing and infrastructural investments. Specifically, the approaches reviewed in this paper, especially online convex optimization by Liu et al. [8][9], adaptive VM consolidation by Beloglazov and Buyya [1], and self-adaptive cloud management by Xu et al. [13] provide an advanced algorithmic basis for implementing the technique.

REFERENCES

- [1] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.
- [2] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.
- [3] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2016.
- [4] Í. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "GreenHadoop: Leveraging green energy in data-processing frameworks," in *Proc. 7th ACM European Conf. Computer Systems (EuroSys)*, 2012, pp. 57–70.
- [5] M. A. Islam, H. Mahmud, S. Ren, and X. Wang, "A carbon-aware incentive mechanism for greening colocation data centers," *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 4–16, 2020.
- [6] A. Khosravi, L. L. H. Andrew, and R. Buyya, "Dynamic VM placement method for minimizing energy and carbon cost in geographically distributed cloud data centers," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 183–196, 2017.
- [7] Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., & Manoharan, P. (2025). Enhancing Retail Sector Customer Engagement Through the Application of Modern and Intelligent AI-Driven Analytical Techniques. *International Journal of Humanities and Information Technology*, 7(01), 19-26.
- [8] Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., Addepalli, P. V., & Bandaru, S. R. (2025). An AI-Driven Classification System for Early Detection of Customer Churn in E-Commerce Platforms. *Journal of Computer Science and Technology Studies*, 7(11), 452-462.
- [9] Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., & Addepalli, P. V. (2025). Customer Sentiment Analysis for Online Shopping Platforms Using NLP and Deep Learning Methods. *Journal of Cyber-Physical Security and Robotics*, 1(02), 1-11.
- [10] Mylavarapu, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., & Kavirayani, U. M. (2025). An Enhanced Intrusion Detection Architecture Using ML-Based Anomaly Detection on Large-Scale Data. *International Journal of Humanities and Information Technology*, 7(01), 20-26.
- [11] Panchali, S. H., Mylavarapu, K. B., Pilli, J., Jannu, P. K., & Mohammad, J. A. (2025). Review of Secure Communication Systems for Cyber-Physical Systems and Industrial IoT. *Journal Of Engineering And Computer Sciences*, 4(6), 349-362.
- [12] Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., Kavirayani, U. M., & Mylavarapu, K. B. (2025). Data-Driven Artificial intelligence (AI)-based System for Identifying Fraudulent Telephony Activities. *Journal of Cyber-Physical Security and Robotics*, 1(02), 23-31.
- [13] Jannu, P. K., Mohammad, J. A., Panchali, S. H., Kavirayani, U. M., Mylavarapu, K. B., & Pilli, J. (2025). A Review of Sensor Data Acquisition Methods for Accurate and Timely Detection of Faults in Mechanical Systems. *Journal of Computer Science and Technology Studies*, 7(9), 796-803.
- [14] Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2025). Emerging Communication Technologies for IoT: 5G and Edge Networking Perspectives. *International Journal of Humanities and Information Technology*, 7(01), 19-28.
- [15] Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., & Pathani, A. K. (2025). Transforming Undergraduate Education Through Modern Pedagogical Innovations. *Journal of Computer Science and Technology Studies*, 7(12), 547-556.
- [16] Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., & Kavuluri, H. V. R. (2025). Big Data and Cloud Computing Integration: A Review of Scalable Information Retrieval Techniques. *Journal of Cyber-Physical Security and Robotics*, 1(02), 12-22.
- [17] Suresh, A. (2025). AI-Driven Business Intelligence Automation: Integrating Data Engineering, Auto-BI, and Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, 6(3), 97-108. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I3P112>
- [18] Suresh, A. (2025). Conversational Analytics Using LLMs: Transforming Enterprise Data Consumption through Natural Language Interfaces. *American International Journal of Computer Science and Technology*, 7(5), 92-102. <https://doi.org/10.63282/3117-5481/AIJCS-T7I5P108>
- [19] Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., & Manoharan, P. (2026). AI-Powered Data Analytics for Strengthening Customer Loyalty and Personalization in Retail. *Journal of Data Analysis and Critical Management*, 2(01), 28-36.
- [20] Manoharan, P., Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., & Shyam, D. (2026). A Review of Digital Tools and Technologies for Enhancing Customer Engagement in Financial Services. *Journal of Contemporary Science and Technology*



- Management*, 2(01), 13-25.
- [21] Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., & Syed, M. H. (2026). Predictive Customer Segmentation in Retail Using Machine Learning and Behavioral Analytics. *European Journal of Applied Science, Engineering and Technology*, 4(2), 268-278.
- [22] Suresh, "Architectural Design of LLM- Powered Conversational BI Systems for High-Volume Enterprise Data Warehouses," 2026 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Bangkok, Thailand, 2026, pp. 1333-1341, doi: 10.1109/ICMLAS67792.2026.11483861.
- [23] Taluri, R. (2024). A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(2), 230-240. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P126>
- [24] Suresh, A. (2024). Auto-BI Frameworks Powered by Generative Artificial Intelligence for Scalable Self-Service Data Analytics in Large Organizations. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(3), 191-201. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P119>
- [25] Taluri, R. (2024). A Cloud-Native Reference Architecture for Data Engineering, Generative AI, and Decision Intelligence Using AWS and Amazon Bedrock. *International Journal of AI, BigData, Computational and Management Studies*, 5(1), 218-227. <https://doi.org/10.63282/3050-9416.IJAIDCMS-V5I1P122>
- [26] Suresh, A. (2025). AI-Driven Business Intelligence Automation: Integrating Data Engineering, Auto-BI, and Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, 6(3), 97-108. <https://doi.org/10.63282/3050-9416.IJAIDCMS-V6I3P112>
- [27] Taluri, R. (2021). Cloud-Native Architectures for Enterprise Financial Data Management, Analytics, and Regulatory Reporting Compliance. *International Journal of Emerging Trends in Computer Science and Information Technology*, 2(2), 101-111. <https://doi.org/10.63282/3050-9246.IJETCSIT-V2I2P112>
- [28] Suresh, A. (2025). Conversational Analytics Using LLMs: Transforming Enterprise Data Consumption through Natural Language Interfaces. *American International Journal of Computer Science and Technology*, 7(5), 92-102. <https://doi.org/10.63282/3117-5481/AIJCSIT-V7I5P108>
- [29] Taluri, R. (2022). Cloud Data Engineering Strategies for Large-Scale Financial Data Integration and Intelligent Corporate Performance Reporting. *International Journal of Emerging Research in Engineering and Technology*, 3(4), 176-188. <https://doi.org/10.63282/3050-922X.IJERET-V3I4P119>
- [30] Jannu, P. K., Mohammad, J. A., Panchali, S. H., Kavirayani, U. M., Mylavarapu, K. B., & Pilli, J. (2026). Recent Advances in Anomaly Identification for IoT Devices Using Machine and Deep Learning Models. *Journal of Data Analysis and Critical Management*, 2(01), 37-47.
- [31] Mylavarapu, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., & Kavirayani, U. M. (2026). Early Diagnosis of Heart Conditions Using AI-Driven on Electronic Health Data. *Journal of Contemporary Science and Technology Management*, 2(01), 26-37.
- [32] Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., Kavirayani, U. M., & Mylavarapu, K. B. (2026). A Survey on Power Optimization Techniques in Contemporary Mobile SoC Architectures. *European Journal of Applied Science, Engineering and Technology*, 4(2), 321-333.
- [33] Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., & Dasari, A. (2026). A Survey on Generative AI Chatbots in Customer Service: Challenges, Benefits, and Use Cases. *Journal of Data Analysis and Critical Management*, 2(01), 17-27.
- [34] Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., & Kavuluri, H. V. R. (2026). Transforming Customer Support with AI: Emerging Tools, Trends and Their Service Effectiveness. *Journal of Contemporary Science and Technology Management*, 2(01), 1-12.
- [35] Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., & Pathani, A. K. (2026). Applications of Machine Learning in Predicting Customer Churn in Telecommunications: A Review Analysis. *European Journal of Applied Science, Engineering and Technology*, 4(2), 257-267.
- [36] Kumar, A., Wadhwa, M., Kalla, D., Konduru, S. C., Nandawat, C., & Sharma, M. (2025, October). Benchmarking the Trade-Offs in Object Detection: Accuracy, Speed, and Energy Efficiency. In *International Conference on Artificial Intelligence and Networking* (pp. 410-422). Cham: Springer Nature Switzerland.
- [37] NR, A. R., Rajasri, T., Praveen, R., Kalla, D., Bendale, S. P., & Venu, N. (2025, April). CAC Training-A Unified Cybersecurity Training Program for Military Staff. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)* (Vol. 3, pp. 569-573). IEEE.
- [38] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, 2(2), 55-62.
- [39] Zeeshan, M., Bhadauria, K., Pahal, L., Nagrath, P., & Kalla, D. (2025, June). Ensemble-Based Deep Learning for Automated Diabetic-Retinopathy Detection Using CNNs and Transfer Learning. In *International Conference on Data Analytics & Management* (pp. 216-228). Cham: Springer Nature Switzerland.
- [40] Aggarwal, A., Agarwal, L., Rella, B. P. R., Nagpal, N., Kalla, D., & Sharma, M. (2025, June). A Performance Comparison of Machine Learning Models for Rain Prediction. In *International Conference on Data Analytics & Management* (pp. 319-328). Cham: Springer Nature Switzerland.
- [41] Nagrath, P., Saini, I., Zeeshan, M., Komal, Komal, & Kalla, D. (2025, June). Predicting Mental Health Disorders with Variational Autoencoders. In *International Conference on Data Analytics & Management* (pp. 38-51). Cham: Springer Nature Switzerland.
- [42] NR, A. R., Rajasri, T., Praveen, R., Kalla, D., Bendale, S. P., & Venu, N. (2025, April). CAC Training-A Unified Cybersecurity Training Program for Military Staff. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)* (Vol. 3, pp. 569-573). IEEE.
- [43] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2024). Hybrid scalable researcher recommendation system using azure data lake analytics. *Journal of Data Analysis and Information Processing*, 12, 76-88.
- [44] Kumar¹, A., Wadhwa¹, M., Kalla, D., Konduru, S. C., Nandawat, C., & Sharma, M. (2026, May). Benchmarking the Trade-Offs in Object Detection: Accuracy, Speed, and Energy. In *Proceedings of International Conference on Artificial Intelligence and Networks: ICAIN 2025, Volume 4* (Vol. 4, p. 410). Springer Nature.
- [45] A. Suresh, "Large Language Models for Explainable and Self-Service Business Intelligence in Large Enterprises," 2026 6th International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2026, pp. 1571-1578, doi: 10.1109/

- ICOECA68095.2026.11485058.
- [46] A. Suresh, "A Comprehensive Study on Auto - BI Systems using Generative AI for Scalable and Explainable Enterprise Analytics," 2026 6th International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2026, pp. 1579-1585, doi: 10.1109/ICOECA68095.2026.11485569.
- [47] Krishnappa, K. H., R. M., & Chavan, L. (2026). ML-Driven Automated Functional Verification Framework for Digital Designs. *Electronics*, 15(12), 2687.
- [48] Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. Available at SSRN 5045491.
- [49] Suresh, "Architectural Design of LLM- Powered Conversational BI Systems for High-Volume Enterprise Data Warehouses," 2026 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Bangkok, Thailand, 2026, pp. 1333-1341, doi: 10.1109/ICMLAS67792.2026.11483861.
- [50] Vadisetty, R., Polamarasetti, A., Varadarajan, V., Kalla, D., & Ramanathan, G. K. (2025, May). Cyber Warfare and AI Agents: Strengthening National Security Against Advanced Persistent Threats (APTs). In International Conference on Intelligence-Based Transformations of Technology and Business Trends (pp. 578-587). Cham: Springer Nature Switzerland.
- [51] J. G. Koomey, "Worldwide electricity used in data centers," *Environmental Research Letters*, vol. 3, no. 3, Art. no. 034008, 2008.
- [52] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Greening geographical load balancing," in *Proc. ACM SIGMETRICS Int. Conf. Measurement and Modeling of Computer Systems*, 2011, pp. 233-244.
- [53] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 62-66, 2011.
- [54] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984-986, 2020.
- [55] A. Shehabi, S. J. Smith, E. Masanet, and J. Koomey, "Data center growth in the United States: Decoupling the demand for services from electricity use," *Environmental Research Letters*, vol. 13, no. 12, Art. no. 124030, 2018.
- [56] M. Xu and R. Buyya, "Managing renewable energy and carbon footprint in multi-cloud computing environments," *Journal of Parallel and Distributed Computing*, vol. 135, pp. 191-202, 2020.
- [57] M. Xu, A. N. Toosi, and R. Buyya, "A self-adaptive approach for managing applications and harnessing renewable energy for sustainable cloud computing," *IEEE Transactions on Sustainable Computing*, advance online publication, 2020.
- [58] 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, 2020.
- [59] Z. Zhou, F. Liu, Y. Xu, R. Zou, H. Xu, J. C. S. Lui, and H. Jin, "Carbon-aware load balancing for geo-distributed cloud services," in *Proc. 2013 IEEE 21st Int. Symp. Modelling, Analysis & Simulation of Computer and Telecommunication Systems*, 2013, pp. 232-241.

