GreenPASS: Zero-Cost Carbon Reduction in the Cloud Through Provider-Assisted Spatial Shifting

Zihan Pan
University of British Columbia
Vancouver, British Columbia, Canada
zihan.pan@ece.ubc.ca

Geoffrey Bian
University of British Columbia
Vancouver, British Columbia, Canada
gbian@student.ubc.ca

Mohammad Shahrad
University of British Columbia
Vancouver, British Columbia, Canada
mshahrad@ece.ubc.ca

Abstract

The global expansion of computing workloads has necessitated exploring ways to reduce emissions. One effective technique is spatial workload shifting, which leverages the varying grid carbon intensity of different geographical regions. However, existing solutions either assume full workload knowledge, making them applicable only to internal workloads, or rely on external middleware. In this paper, we explore how spatial workload shifting can be performed cooperatively between users and providers in the cloud context. We demonstrate how this capability can be offered as a service without additional cost to users, by means of providers harnessing grid price variations to cover added costs. We explore various design elements to build this architecture and identify its potential.

CCS Concepts

• **Computer systems organization → Cloud computing**; • **Social and professional topics → Sustainability**.

Keywords

Cloud Computing, Data Centers, Sustainability, Geospatial Shifting

ACM Reference Format:
Zihan Pan, Geoffrey Bian, and Mohammad Shahrad. 2026. GreenPASS: Zero-Cost Carbon Reduction in the Cloud Through Provider-Assisted Spatial Shifting. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3744255.3811732

1 Introduction

The unprecedented growth in data center energy consumption has stressed electrical grids and driven substantial operational emissions [5, 30]. Temporal and spatial workload shifting have both been explored to mitigate these emissions by moving workloads to cleaner times or locations, respectively. Among these strategies, spatial shifting often yields much larger emission reductions as carbon intensity varies dramatically across regions [59].

Hyperscalers have pursued spatial shifting for their internal workloads for several years. For example, as early as 2020, Google began reducing the emissions of its services through carbon-aware resource scheduling [15]. Several prior studies implicitly assume full visibility into workload characteristics, as they do not address the challenges posed by third-party applications in spatial shifting [36].

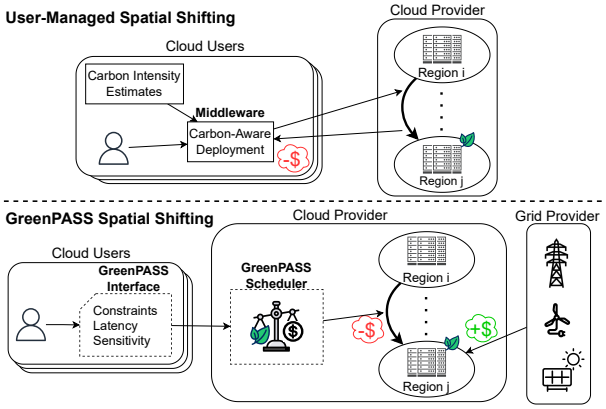

Figure 1: GreenPASS eliminates the cost and complexity of carbon-aware deployment middleware for cloud users while offsetting provider-side shifting costs through grid savings.

In practice, however, spatial shifting of third-party workloads in the cloud is extremely challenging. Cloud workloads are typically opaque to the providers; they have little insight into application logic and operational constraints, especially privacy and compliance requirements. This makes spatial shifting difficult for providers serving diverse users. Consequently, carbon-aware spatial shifting for cloud users has evolved into a patchwork of domain-specific solutions tailored to web services [37], AI inferences [33], serverless applications [24], etc. The overhead of operating and maintaining such middleware is discouraging for users. More importantly, limited access to accurate information about a host data center's energy mix forces spatial shifting middleware to rely on coarse-grained carbon-intensity metrics (often at ISO-level) reported by platforms like Electricity Maps [39] and WattTime [63]. In contrast, providers possess precise knowledge of how much power comes from the ISO grid, on-site storage [16], cogeneration [55], and other sources.

We advocate for a third approach, involving cooperation between the user and provider (Fig. 1). We envisage that users seeking carbon reduction through spatial workload shifting could receive it for free as a service from the provider. No need to manage the middleware, no need to pay for egress costs, and no need to pay for any data duplication or compute time overlap. They provide information on their possible host regions and degree of latency tolerance, and ensure that the application itself has region-agnostic bindings or variants. This service is best-effort, however; the provider would do it by harnessing the grid price differences across current and target regions; only when the marginal electricity price gains can at least cover the associated shifting costs, would it do so. We refer to this architecture as Green Provider-Assisted Spatial Shifting

E-Energy '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2011-6/26/06
https://doi.org/10.1145/3744255.3811732

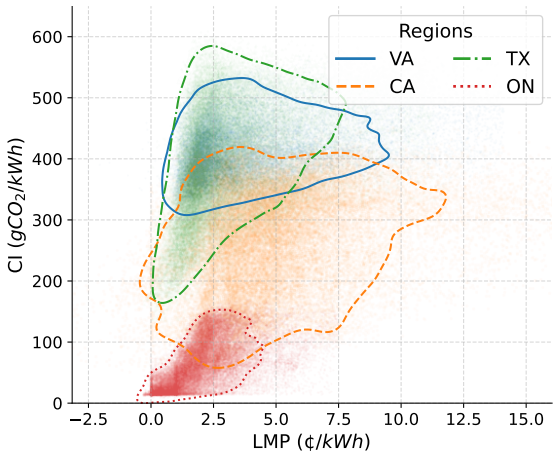

Figure 2: Carbon Intensity (CI) and Locational Marginal Price (LMP) in four North American regions from 2020 to 2024.

(GreenPASS). In this vision paper, we explore GreenPASS's potential by answering the following research questions:

(1) If we consider common regions in a geographical boundary (in this paper, North America), can the degree of price variation indeed cover the costs of spatial shifting for providers?
(2) Even with covered shifting costs, are emissions reductions meaningful and sustained over time?
(3) How do emissions reductions differ across data access patterns?
(4) What strategies can be used to maximize carbon reduction?

By addressing these questions, this paper seeks to deepen understanding of spatial shifting for third-party applications and to pave the way for broader adoption of provider-assisted spatial shifting.

2 Background & Motivation

2.1 Carbon Intensity and Electricity Price

Carbon intensity (CI) measures the amount of greenhouse gas emissions per unit of energy produced, measured in gCO_2eq/kWh [39], which varies over time and across regions due to the generation mix of local grids. Prior work has leveraged this heterogeneity through techniques such as workload shifting [7, 24, 37, 56, 59].

Locational marginal price (LMP) is the wholesale price of electricity at a given time and location [57], driven by supply-demand dynamics. Variations in LMP are reflected in data centers' energy cost, as market-based electricity rate indexed to LMP is prevalent in the energy retail market [9, 12, 45, 46], and some data centers purchase energy directly from the wholesale market. Prior work has utilized this variation to cut data center energy costs [31, 49, 50].

Fig. 2 shows a scatter plot of the hourly-averaged CI and LMP in four North American regions (Virginia, Texas, California, and Ontario) from 2020 to 2024 (outliers removed), with the enclosed area covering 90% of the data points in each region. The high degree of variation and distinct patterns across regions highlight the potential of workload shifting. Ontario exhibits relatively lower carbon intensity, while California shows the highest variability in LMP. Notably, negative LMP occurs when energy demand cannot absorb excess supply while some generators are unable to reduce output, creating attractive opportunities for workload shifting.

The variation in CI and LMP across time and regions creates co-optimization opportunities to execute workloads with both lower carbon footprint and reduced energy costs. Prior work has investigated this co-optimization [14, 25, 27, 40, 41, 67, 69], while existing literature has yet to explore this in a cloud context, specifically regarding the collaboration between cloud providers and users.

2.2 Carbon-Aware Spatial Shifting

Hyperscalers have constructed globally distributed infrastructures, enabling the exploitation of CI variations across the world [17, 52]. As an effective carbon reduction technique, spatial workload shifting has received increasing attentions recently [37, 56, 65].

Although users can deploy middleware for spatial shifting, doing so increases operational complexity. Shifting workloads can also incur cross-region data transfer egress fees and additional storage costs due to duplication; when charged at list prices, these can tangibly increase user bills. Finally, user-managed middleware cannot leverage provider-side grid-level signals such as LMP for holistic optimization.

Conversely, workload shifting is more feasible and efficient from the cloud providers' perspective. Economically, provider-side operational costs for data transmission and storage is often lower than the listed user prices [10], making spatial shifting significantly cheaper at scale. Moreover, providers have greater visibility into the grid through their wholesale contracts, enabling them to take advantage of various programs (e.g., demand response plans). Several prior works have exploited this capability [8, 60, 68].

In practice, providers may not be able to shift user workloads due to major implications on performance, data access patterns, and regulatory compliance [26]. Shifting without users' guidance may violate performance and/or regulatory requirements [24]. Besides, the monetary cost of shifting conflicts with profit-maximization objectives unless adequate incentives are provided to offset these shifting costs. We believe enabling effective carbon-aware spatial shifting requires a cooperative approach to bridge the gap between user constraints, provider capabilities, and provider profitability.

3 GreenPASS

To facilitate carbon-aware spatial shifting for cloud workloads, we propose a collaborative framework in which 1) users provide workload characteristics, constraints, and their willingness to reduce carbon at no additional cost, and 2) providers shift workloads according to user-defined constraints and leverage LMP variations across regions from the grid to cover the cost of spatial shifting. Under this framework, both cloud users and providers can reduce carbon emissions without incurring additional monetary cost for either party, aligning mutual benefit with environmental objectives.

3.1 User Interface

In our vision, cloud users interested in GreenPASS need to provide five categories of information:

(1) **Residency Constraints**: Without residency constraints, automated shifting may inadvertently violate data sovereignty regulations (e.g., HIPAA [2], GDPR [47]). Users must therefore

define the operational boundary either through an explicit region *allow-list* or global consent. Spatial shifting remains transparent to the user's billing model by continuing to maintain the original region's metadata for each resource, now treated as the *home* region.

(2) **State Handling**: Spatial shifting for stateless applications is highly efficient, as the one-time overhead of transferring container or VM images gets amortized by the carbon savings over time. In contrast, stateful applications necessitate specialized migration primitives, such as checkpoint-restore or global state synchronization, to prevent data divergence. Users must verify that their workload supports such transitions without breaking functionality; either implicitly through participation terms or explicitly, such as by selecting a checkbox during service setup.

(3) **Data Access Policy**: Cloud workloads often depend on remote data dependencies, both for I/O operations and external API calls. For write-heavy workloads requiring strict consistency, a centralized data model may simplify application logic but risks significant performance degradation due to increased cross-region Round Trip Time (RTT) and bandwidth constraints. In contrast, read-heavy workloads can be optimized through regional replication of static assets, such as model weights, to maintain high throughput. Users should communicate to the provider whether specific storage media (e.g., Redis data store) should be part of the migration payload or if the application will manage its own remote data retrieval.

(4) **Latency Sensitivity**: Emission reduction can be maximized when applications are not latency-sensitive. However, many real-world, user-facing cloud services operate under strict latency requirements. To account for this, users can explicitly communicate an application's latency tolerance, by specifying an absolute tail-latency cap (e.g., at the 95th percentile) or by allowing a bounded relative increase (e.g., 10–15%) in average response time compared to full deployment in the home region.

(5) **Alternative Resources**: Users can increase the likelihood of a successful carbon-aware migration by defining *resource flexibility*. This involves specifying a range of acceptable VM sizes and container types, or even providing cross-architecture support (e.g., allowing ARM as an alternative to x86). This flexibility ensures that the workload can be placed on the cleanest available region, even if the primary resource type is unavailable.

3.2 Cost- and Delay-Aware Carbon Reduction

We abstract each user workload, together with its characteristics and constraints, as a job (j), which is the basic unit of scheduling. GreenPASS estimates the effect of shifting j from its home region (s_j) to each candidate region (d_j) along on three metrics: 1) **latency** due to remote data access overheads, 2) **monetary cost** of shifting operations and savings from LMP variations, and 3) **carbon emissions** of shifting operations and reductions from CI variation.

To prevent Service Level Agreement (SLA) violations, the provider must ensure that the latency due to shifting from the home region s_j to the destination region d_j ($T^{\text{shift}}_{j,s_j\to d_j}$) remains within the users' tolerance. As mentioned in §3.1, users can specify latency slack in various ways, including setting a ratio τ of job execution time (T^{exec}_j):

$$T^{\text{shift}}_{j,s_j\to d_j} \le \tau \times T^{\text{exec}}_j \tag{1}$$

The provider also ensures that the operational costs of shifting ($\text{Cost}^{\text{shift}}_{j,s_j\to d_j}$) are offset by energy cost savings. We apply different optimization scopes to balance the costs of data transmission and duplication against the savings from LMP variation:

1) **Per-job**, where a job is shifted if its energy cost savings can cover its shifting costs:

$$\text{Cost}^{\text{shift}}_{j,s_j\to d_j} \le (\text{Cost}^{\text{energy}}_{j,s_j} - \text{Cost}^{\text{energy}}_{j,d_j}) \tag{2}$$

2) **Per-user**, where each user maintains a budget pool that accumulates savings from prior shifted jobs for later reuse:

$$\sum_{j\in J_{\text{user}}} \text{Cost}^{\text{shift}}_{j,s_j\to d_j} \le \sum_{j\in J_{\text{user}}} (\text{Cost}^{\text{energy}}_{j,s_j} - \text{Cost}^{\text{energy}}_{j,d_j}) \tag{3}$$

3) **Global**, which extends the budget pool from individuals to all participating users, maximizing the system-wide shifting potential:

$$\sum_{j\in J_{\text{all}}} \text{Cost}^{\text{shift}}_{j,s_j\to d_j} \le \sum_{j\in J_{\text{all}}} (\text{Cost}^{\text{energy}}_{j,s_j} - \text{Cost}^{\text{energy}}_{j,d_j}) \tag{4}$$

Subject to these constraints, GreenPASS selects the region(s) with maximum carbon saving (where $\text{Carbon}^{\text{shift}}_{j,s_j\to s_j} = 0$):

$$\arg\max_{d_j} \left[\text{Carbon}^{\text{exec}}_{j,s_j} - \text{Carbon}^{\text{exec}}_{j,d_j} - \text{Carbon}^{\text{shift}}_{j,s_j\to d_j}\right] \tag{5}$$

Therefore, the provider deploys each job to the region that minimize carbon while satisfying both latency and cost constraints.

4 Evaluation

4.1 Methodology

We chose a simulation-based evaluation methodology that enables testing at a cluster scale under various design choices, workload constraints, and cost assumptions. We evaluate GreenPASS in terms of operational carbon reduction since it manages workload deployment over existing data center capacity. Embodied carbon should not influence scheduling decisions since they are sunk carbon as shown by prior work [6]. We use production traces from the Alibaba 2020 GPU trace dataset [64] as our workload. This dataset contains user job requests from a Machine-Learning-as-a-Service (MLaaS) cluster over two months, from which we take a one-week subset (~70K jobs) starting from the second month. Jobs in the first month are used for collecting historical information for prediction. We choose this dataset because it contains a mix of training and inference jobs, capturing the heterogeneity of cloud workloads in terms of data and latency requirements. To the best of our knowledge, it is the only publicly available trace with comprehensive information on the utilization, duration, and data usage of job resources, which is necessary for a holistic assessment of GreenPASS.

We simulate four North American regions: Virginia, California, Texas, and Ontario, approximately corresponding to Google Cloud regions us-east4, us-west2, us-south1, and northamerica-northeast2 [17]. We construct an evenly mixed mapping by distributing users in the trace across the four regions (hereafter denoted as *Mixed*). Unless stated otherwise, results in the following sections are reported for this Mixed setting, with metrics averaged over three random user assignments.

Similar to prior work [4, 24, 54, 62], we estimate energy consumption with a linear model. We obtain CI and LMP data from prior work artifacts [25, 59] collected from Electricity Maps [39] and Grid Status [57]. We used instance prices from Google Cloud to represent user computation costs, and their listed prices for data egress and storage to represent user-side shifting costs when users shift workloads with their own middleware [3, 18, 20, 22, 23]. On the provider side, the exact unit costs of data transmission and storage are unknown to us, but we assume these costs should be no higher than the listed prices, otherwise it would not be profitable to host such services. We therefore model them with a price range bounded by optimistic estimates from prior work ($R_{\min}$) as the lower bound, and the listed prices ($R_{\max}$) as the upper bound. Specifically, two reports have provided optimal unit bandwidth and storage cost estimates at data center scale [53, 61], which we adopt as our lower bound. We interpolate between these bounds by a provider shifting cost factor $\alpha \in [0, 1]$:

$$R_{\text{provider}} = \alpha \times R_{\max} + (1 - \alpha) \times R_{\min} \tag{6}$$

A smaller α assumes lower provider-side shifting cost, while a larger α is more conservative. We also evaluate a range of remote data access patterns that affect the amount of data transmission and duplication of shifting. Detailed formulations for latency, cost, and carbon in our simulation are provided in Appendix A.

We evaluate our framework in both oracle and prediction settings for future job characteristics, CI, and LMP. We estimate jobs' duration and power consumption based on historical jobs from the trace, and we predict CI and LMP with LightGBM [29]. Details of our prediction methods are provided in Appendix B. Nevertheless, prediction is not the focus of this paper, and recent work has proposed more sophisticated CI forecasting methods [28, 32, 66].

As the aim of our evaluation is to estimate the feasibility of GreenPASS, we adopt a few simplifying assumptions: 1) we assume no provider capacity limit since we do not have access to such information; 2) we do not model complex dynamics of supply-demand across regions affecting resource prices; 3) we use the average rather than the marginal carbon intensity, assuming workload shifting is insufficient to alter the marginal grid dispatch, which is a complex bi-directional interaction explored in prior studies [34, 35, 58]. Incorporating data center capacity constraints, cloud price variations, and marginal grid effects is left for future work.

4.2 Effectiveness of GreenPASS

Fig. 3 compares user-managed shifting using custom middleware with provider-assisted shifting using GreenPASS under different latency constraints, assuming the optimal provider shifting cost factor ($\alpha = 0$). Operational carbon is reduced as users tolerate higher latency, but eventually saturates as transmission and storage emissions offset the benefits of shifting. The following sub-sections use latency tolerance of 1%. In user-managed shifting, users incur additional costs for data transmission and storage duplication at listed prices, which could cause up to 4.5% cost increases comparing with simply running the jobs in their home regions. Cloud providers also offer lower-cost compute options such as spot VMs. As spot prices fluctuate over time, we approximate their cost as 50% of the on-demand price in our analysis [20]. Under this pricing model, the relative overhead of user-managed shifting becomes even more

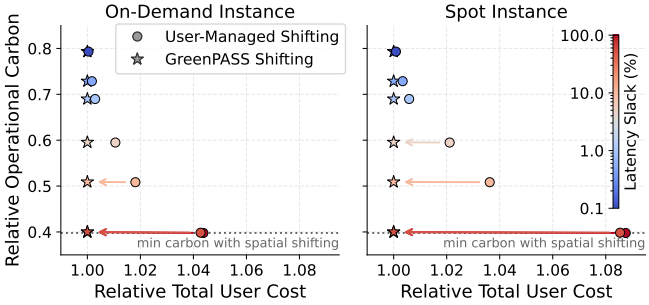

Figure 3: GreenPASS allows reducing emissions without adding cost to users. With lower computation cost, e.g., spot VMs (right), this relative cost reduction is greater.

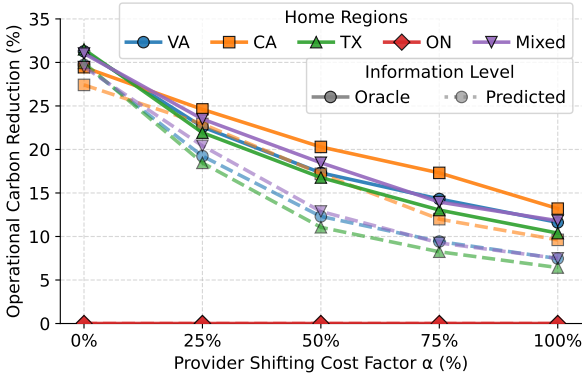

Figure 4: Carbon reduction potential varies by home region.

significant, increasing total costs by up to 9%. Compared with user-managed shifting, GreenPASS eliminates these user-incurred costs while achieving the same carbon reduction when the provider operates data transfer and storage at the lower-bound cost.

4.3 Region and Provider Shifting Cost

Fig. 4 shows the sensitivity of carbon reduction to provider cost factor under both oracle and prediction settings. As expected, a higher provider cost factor limits carbon reduction since fewer jobs can be shifted under the fixed LMP price differential. Solid lines show oracle-based decisions, while dashed lines use predicted job characteristic, CI, and LMP values. Carbon reductions for various home regions are shown, as well. As Ontario has the lowest CI most of the time, designating it as the home region leaves no incentive to shift jobs outward. Conversely, LMP in California has higher variability, creating greater potential for energy cost savings and thus enabling higher carbon reduction as provider cost factor increases.

4.4 Data Access Patterns

We discussed how users can specify data access policies in §3.1. Here, we investigate three specific policies: 1) **Remote I/O**, where all data is accessed remotely, from the home regions; 2) **Duplication**, where data is transferred to the destination region before execution and stored until the job completes; 3) **Duplication & Reuse**, where, additionally, data is reused across recurring jobs deployed in the same region. A detailed description on how these

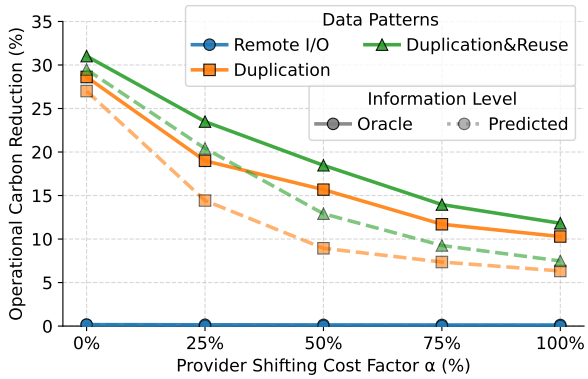

Figure 5: Duplicating and reusing data in the destination region yields the most carbon savings.

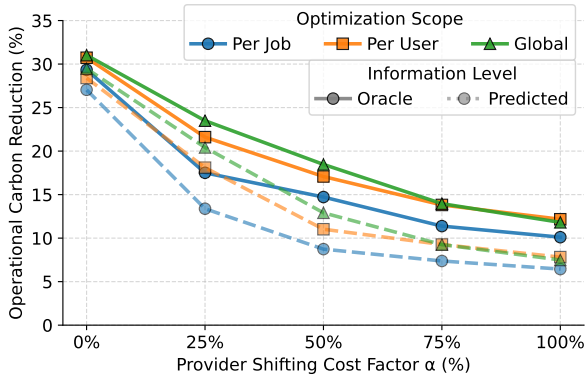

Figure 6: Expanding the optimization scope from jobs to users to global enables more carbon savings.

patterns affect shifting overhead is provided in Appendix A.2. Fig. 5 shows the impact of these patterns on carbon reduction potential. Remote I/O yields negligible carbon reduction because most jobs in our traces are I/O-intensive, incurring substantial cross-region round-trip latencies; as a result, only a small fraction can be shifted without violating latency slack. Duplicating the data at the destination region enables local data access during execution, ensuring tolerable latency. Furthermore, data reuse across recurring jobs reduces the total data transfer volume, leading to lower shifting cost and higher overall carbon reduction.

4.5 Optimization Scope

Fig. 6 shows the carbon reduction under different optimization scopes. One benefit of GreenPASS is the ability to jointly optimize workloads from multiple users. Making decisions at a larger scale allows energy savings from some jobs to offset the shifting costs of others, enabling greater overall carbon reduction. However, a limitation of greedy budget allocation is that jobs with a high data-to-compute ratio would be shifted as soon as they arrive, leaving insufficient budget for future jobs that could achieve higher carbon reduction with lower data usage.

4.6 Temporal Sensitivity

Earlier results focused on the first week of August 2020, aligning with Alibaba workload trace's original collection window. To evaluate the robustness of GreenPASS over time, we simulate the same

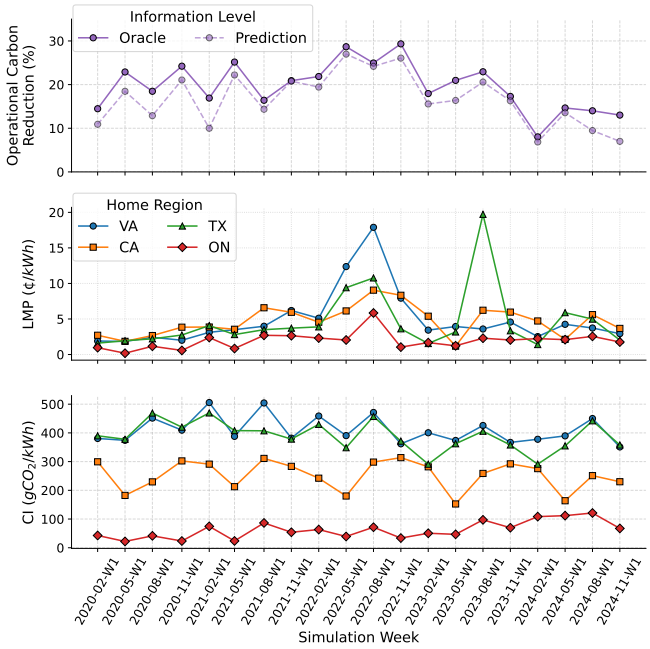

Figure 7: Operational carbon reduction from GreenPASS is relatively consistent throughout five years.

workload starting from the second month of each quarter over five years to capture seasonal and annual variation in CI and LMP. Results are shown in Fig. 7, where each marker point represents the carbon reduction (top), average LMP (middle), or average CI (bottom) of the first week of the month. Overall, non-negligible carbon reduction is observed in all tested periods, with variations driven by temporal patterns in CI and LMP across regions. Periods with high regional LMP disparities (e.g., Virginia in August 2022 and Texas in August 2023) increased energy cost-saving potential. This allows the optimizer to cover shifting cost more easily, giving slightly higher carbon reduction with both oracle and predicted information. In contrast, 2024 achieves lower reductions due to smaller LMP differences and higher average Ontario CI.

5 Conclusion

We propose a collaborative model for reducing emissions from third-party cloud applications by enabling users to delegate spatial workload shifting to the provider through a lightweight interface. By exploiting inter-region LMP differences, providers can offset shifting costs and offer carbon-aware placement transparently and at no additional cost to users. We evaluate the architecture's feasibility and the key factors governing its effectiveness.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was in part supported by Rogers Communications Canada Inc. and Natural Sciences and Engineering Research Council of Canada (NSERC). Computational resources from the Digital Research Alliance of Canada (DRAC) enabled our studies.

References

[1] AMD. 2026. *5th Generation AMD EPYC™ Server CPUs.* Retrieved January 26, 2026 from https://www.amd.com/en/products/processors/server/epyc/9005-series.html

[2] Brian K. Atchinson and Daniel M. Fox. 1997. The politics of the Health Insurance Portability and Accountability Act. *Health Affairs (Millwood)* 16, 3 (1997), 146–150.

[3] AWS. 2026. *AWS Fargate Pricing.* Retrieved January 26, 2026 from https://aws.amazon.com/fargate/pricing/

[4] Hanan Awwad, Changyuan Lin, Rabab Ward, and Mohammad Shahrad. 2025. Estimating the Carbon Footprint of Serverless Functions on a Public Cloud Platform. In *Proceedings of the 3rd Workshop on SErverless Systems, Applications and MEthodologies* (Rotterdam, Netherlands) *(SESAME' 25)*. ACM, 12–20.

[5] Seth Ayers, Sara Ballan, Vanessa Gray, and Rosie McDonald. 2024. *Measuring the Emissions and Energy Footprint of the ICT Sector: Implications for Climate Action.* Technical Report.

[6] Noman Bashir, Varun Gohil, Anagha Belavadi Subramanya, Mohammad Shahrad, David Irwin, Elsa Olivetti, and Christina Delimitrou. 2024. The Sunk Carbon Fallacy: Rethinking Carbon Footprint Metrics for Effective Carbon-Aware Scheduling. In *Proceedings of the 2024 ACM Symposium on Cloud Computing* (Redmond, WA, USA) *(SoCC '24)*. ACM, 542–551.

[7] Mohak Chadha, Thandayuthapani Subramanian, Eishi Arima, Michael Gerndt, Martin Schulz, and Osama Abboud. 2023. GreenCourier: Carbon-Aware Scheduling for Serverless Functions. In *Proceedings of the 9th International Workshop on Serverless Computing* (Bologna, Italy) *(WoSC '23)*. ACM, 18–23.

[8] Hao Chen, Yijia Zhang, Michael C. Caramanis, and Ayse K. Coskun. 2019. EnergyQARE: QoS-Aware Data Center Participation in Smart Grid Regulation Service Reserve Provision. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 4, 1, Article 2 (Jan. 2019), 31 pages.

[9] Federal Energy Regulatory Commission. 2023. *Electric Market-Based Rates.* Retrieved January 26, 2026 from https://www.ferc.gov/power-sales-and-markets/electric-market-based-rates

[10] Competition and Markets Authority. 2024. *Cloud services market investigation: Egress fees working paper.* Technical Report. Competition and Markets Authority, London, UK. https://assets.publishing.service.gov.uk/media/664f2556993111924d9d3aa8/240521_-_Egress_Fees__.pdf

[11] Benjamin Davy. 2021. *Estimating AWS EC2 Instances Power Consumption.* Retrieved January 26, 2026 from https://medium.com/teads-engineering/estimating-aws-ec2-instances-power-consumption-c9745e347959

[12] Virginia Electric and Power Company. 2025. Schedule MBR Large General Service Market-Based Rate. https://cdn-dominionenergy-prd-001.azureedge.net/-/media/content/rates-and-tariffs/pdfs/virginia/mbr.pdf

[13] AT&T Center for Virtualization at Southern Methodist University. 2023. *Google Cloud Inter-Region Latency and Throughput.* Retrieved January 26, 2026 from https://lookerstudio.google.com/u/0/reporting/fc733b10-9744-4a72-a502-92290f608571/page/70YCB

[14] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. 2012. It's not easy being green. In *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication* (Helsinki, Finland) *(SIGCOMM '12)*. ACM, 211–222.

[15] Google. 2020. *Our data centers now work harder when the sun shines and wind blows.* Retrieved January 26, 2026 from https://blog.google/innovation-and-ai/infrastructure-and-cloud/global-network/data-centers-work-harder-sun-shines-wind-blows/

[16] Google. 2025. *How we got to 100 million cells in our global Li-ion rack battery fleet.* Retrieved January 26, 2026 from https://cloud.google.com/blog/topics/systems/100-million-li-ion-cells-in-google-data-centers

[17] Google. 2026. *Cloud locations.* Retrieved January 26, 2026 from https://cloud.google.com/about/locations

[18] Google. 2026. *Cloud Storage pricing.* Retrieved January 26, 2026 from https://cloud.google.com/storage/pricing

[19] Google. 2026. *Power Usage Effectiveness.* Retrieved January 26, 2026 from https://datacenters.google/efficiency/

[20] Google. 2026. *Spot VMs pricing.* Retrieved January 26, 2026 from https://cloud.google.com/spot-vms/pricing

[21] Google. 2026. *Virtual Machines Pricing General-purpose machine type family.* Retrieved May 9, 2026 from https://cloud.google.com/products/compute/pricing

[22] Google. 2026. *Virtual Private Cloud pricing.* Retrieved January 26, 2026 from https://cloud.google.com/vpc/pricing

[23] Google. 2026. *VM instance pricing.* Retrieved January 26, 2026 from https://cloud.google.com/compute/vm-instance-pricing

[24] Viktor Urban Gsteiger, Pin Hong (Daniel) Long, Yiran (Jerry) Sun, Parshan Javanrood, and Mohammad Shahrad. 2024. Caribou: Fine-Grained Geospatial Shifting of Serverless Applications for Sustainability. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles* (Austin, TX, USA) *(SOSP '24)*. ACM, 403–420.

[25] Antonio Guillen, Avisek Naug, Vineet Gundecha, Sahand Ghorbanpour, Ricardo Luna Gutierrez, Ashwin Ramesh Babu, Munther Salim, Shubhanker Banerjee, Eoin H. Oude Essink, Damien Fay, and Soumyendu Sarkar. 2025. SustainCluster: Benchmarking Dynamic Multi-Objective Geo-Distributed Workload Management for Sustainable Data Center Cluster. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track.* Under Review – Details to be updated upon publication.

[26] Praveen Gupta, Arshia Moghimi, Devam Sisodraker, Mohammad Shahrad, and Aastha Mehta. 2025. Growlithe: A Developer-Centric Compliance Tool for Serverless Applications. In *2025 IEEE Symposium on Security and Privacy (SP)*. 3161–3179.

[27] Mohammad A. Islam, Hasan Mahmud, Shaolei Ren, and Xiaorui Wang. 2020. A Carbon-Aware Incentive Mechanism for Greening Colocation Data Centers. *IEEE Transactions on Cloud Computing* 8, 1 (2020), 4–16.

[28] Mathew Joseph, Tanush Savadi, and Abel Souza. 2025. LiteCast: A Lightweight Forecaster for Carbon Optimizations. arXiv:2511.06187 [cs.DC] https://arxiv.org/abs/2511.06187

[29] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

[30] Bran Knowles. 2021. *ACM TechBrief: Computing and Climate Change.* Technical Report.

[31] Kien Le, Ricardo Bianchini, Jingru Zhang, Yogesh Jaluria, Jiandong Meng, and Thu D. Nguyen. 2011. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* (Seattle, Washington) *(SC '11)*. ACM, Article 22, 12 pages.

[32] Amy Li, Sihang Liu, and Yi Ding. 2025. Uncertainty-Aware Decarbonization for Datacenters. *SIGENERGY Energy Inform. Rev.* 4, 5 (April 2025), 141–147.

[33] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. 2024. Towards Environmentally Equitable AI via Geographical Load Balancing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (Singapore, Singapore) *(e-Energy '24)*. ACM, 291–307.

[34] Liuzixuan Lin and Andrew A Chien. 2023. Adapting Datacenter Capacity for Greener Datacenters and Grid. In *Proceedings of the 14th ACM International Conference on Future Energy Systems* (Orlando, FL, USA) *(e-Energy '23)*. ACM, 200–213.

[35] Liuzixuan Lin, Victor M. Zavala, and Andrew A. Chien. 2021. Evaluating Coupling Models for Cloud Datacenters and Power Grids. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems* (Virtual Event, Italy) *(e-Energy '21)*. ACM, 171–184.

[36] Julia Lindberg, Yasmine Abdennadher, Jiaqi Chen, Bernard C. Lesieutre, and Line Roald. 2021. A Guide to Reducing Carbon Emissions through Data Center Geographical Load Shifting. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems* (Virtual Event, Italy) *(e-Energy '21)*. ACM, 430–436.

[37] Diptyaroop Maji, Ben Pfaff, Vipin P R, Rajagopal Sreenivasan, Victor Firoiu, Sreeram Iyer, Colleen Josephson, Zhelong Pan, and Ramesh K Sitaraman. 2023. Bringing Carbon Awareness to Multi-cloud Application Delivery. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems* (Boston, MA, USA) *(HotCarbon '23)*. ACM, Article 6, 6 pages.

[38] Electricity Maps. 2025. *New 72-hour grid forecasts: Advanced load optimization for greater carbon and cost savings.* Retrieved January 26, 2026 from https://www.electricitymaps.com/technology/new-72-hour-grid-forecasts

[39] Electricity Maps. 2026. *Carbon Intensity.* Retrieved January 26, 2026 from https://app.electricitymaps.com/developer-hub/api/signals#carbon-intensity

[40] Jorge Murillo, Walid A. Hanafy, David Irwin, Ramesh Sitaraman, and Prashant Shenoy. 2024. CDN-Shifter: Leveraging Spatial Workload Shifting to Decarbonize Content Delivery Networks. In *Proceedings of the 2024 ACM Symposium on Cloud Computing* (Redmond, WA, USA) *(SoCC '24)*. ACM, 505–521.

[41] Avisek Naug, Antonio Guillen, Ricardo Luna, Vineet Gundecha, Cullen Bash, Sahand Ghorbanpour, Sajad Mousavi, Ashwin Ramesh Babu, Dejan Markovikj, Lekhapriya D Kashyap, Desik Rengarajan, and Soumyendu Sarkar. 2024. SustainDC: benchmarking for sustainable data center control. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 3192, 40 pages.

[42] Nvidia. 2016. *NVIDIA Tesla P100 GPU Accelerator.* Technical Report. https://images.nvidia.com/content/tesla/pdf/nvidia-tesla-p100-PCIe-datasheet.pdf

[43] Nvidia. 2020. *NVIDIA T4 70W Low Profile PCIe GPU Accelerator.* Technical Report. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf

[44] Nvidia. 2020. *NVIDIA V100 Tensor Core GPU.* Technical Report. https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf

[45] Electric Reliability Council of Texas. 2026. *Market Prices*. Retrieved January 26, 2026 from https://www.ercot.com/mktinfo/prices

[46] Independent Electricity System Operator. 2026. *Pricing for Medium and Large Businesses*. Retrieved January 26, 2026 from https://www.ieso.ca/Learn/Electricity-Pricing-Explained/Medium-and-Large-Businesses

[47] The European Parliament and The Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[48] PJM. 2026. *Day-Ahead Hourly LMPs*. Retrieved January 26, 2026 from https://dataminer2.pjm.com/feed/da_hrl_lmps

[49] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. 2009. Cutting the electric bill for internet-scale systems. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication* (Barcelona, Spain) *(SIGCOMM '09)*. ACM, 123–134.

[50] Lei Rao, Xue Liu, Le Xie, and Wenyu Liu. 2010. Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi-Electricity-Market Environment. In *2010 Proceedings IEEE INFOCOM*. 1–9.

[51] Nick Sciarrilli. 2019. AWS 100Gb Network Performance. https://github.com/sciarrilli/aws_network_performance GitHub repository, accessed January 2026.

[52] Amazon Web Services. 2026. *Regions and Availability Zones*. Retrieved January 26, 2026 from https://aws.amazon.com/about-aws/global-infrastructure/regions_az/?nc1=h_ls

[53] Dave Sierra. 2025. *The Economics of Exabyte Data Storage: A 10-year TCO analysis of Solidigm™ QLC SSDs with VAST Data vs. traditional HDDs with CEPH*. Technical Report. Solidigm. https://www.solidigm.com/products/technology/economics-of-exabyte-data-storage.html

[54] Thibault Simon, David Ekchajzer, Adrien Berthelot, Eric Fourboul, Samuel Rince, and Romain Rouvoy. 2025. BoaviztAPI: A Bottom-Up Model to Assess the Environmental Impacts of Cloud Services. *SIGENERGY Energy Inform. Rev.* 4, 5 (April 2025), 84–90.

[55] Brian M Smith, Palash Kumar Bhowmik, Ramon Ken Yoshiura, Eric T Whiting, Matthew William Anderson, Matthew R Sgambati, and Kara G Cafferty. 2025. Accelerating Nuclear-Integrated Data Center Pursuits in the USA: SWOT Analysis, Power-Thermal Management Strategies and Demonstration Plan. *Nuclear Engineering and Design* 446, INL/JOU-25-83648 (2025).

[56] Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2024. CASPER: Carbon-Aware Scheduling and Provisioning for Distributed Web Services. In *Proceedings of the 14th International Green and Sustainable Computing Conference* (Toronto, ON, Canada) *(IGSC '23)*. ACM, 67–73.

[57] Grid Status. 2025. *LMP Dataset Overviews*. Retrieved January 26, 2026 from https://docs.gridstatus.io/data-guides

[58] Thanathorn Sukprasert, Noman Bashir, Abel Souza, David Irwin, and Prashant Shenoy. 2024. On the Implications of Choosing Average versus Marginal Carbon Intensity Signals on Carbon-aware Optimizations. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (Singapore, Singapore) *(e-Energy '24)*. ACM, 422–427.

[59] Thanathorn Sukprasert, Abel Souza, Noman Bashir, David Irwin, and Prashant Shenoy. 2024. On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud. In *Proceedings of the Nineteenth European Conference on Computer Systems* (Athens, Greece) *(EuroSys '24)*. ACM, 924–941.

[60] Qihang Sun, Shaolei Ren, Chuan Wu, and Zongpeng Li. 2016. An online incentive mechanism for emergency demand response in geo-distributed colocation data centers. In *Proceedings of the Seventh International Conference on Future Energy Systems* (Waterloo, Ontario, Canada) *(e-Energy '16)*. ACM, Article 3, 13 pages.

[61] TeleGeography. 2025. *Executive Summary: IP Networks Research Service*. Technical Report. PriMetrica, Inc. https://www2.telegeography.com/hubfs/assets/product-tear-sheets/product-page-content-samples/global-internet-geography/telegeography-global-internet-geography-executive-summary.pdf

[62] Bogdan Marius Tudor and Yong Meng Teo. 2013. On understanding the energy consumption of ARM-based multicore servers. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems* (Pittsburgh, PA, USA) *(SIGMETRICS '13)*. ACM, 267–278.

[63] WattTime. 2026. *Data Signals*. Retrieved January 26, 2026 from https://watttime.org/data-science/data-signals/

[64] Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. 2022. MLaaS in the Wild: Workload Analysis and Scheduling in Large-Scale Heterogeneous GPU Clusters. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. USENIX Association, Renton, WA, 945–960. https://www.usenix.org/conference/nsdi22/presentation/weng

[65] Li Wu, Walid A. Hanafy, Abel Souza, Khai Nguyen, Jan Harkes, David Irwin, Mahadev Satyanarayanan, and Prashant Shenoy. 2025. CarbonEdge: Leveraging Mesoscale Spatial Carbon-Intensity Variations for Low Carbon Edge Computing. In *Proceedings of the 34th International Symposium on High-Performance Parallel and Distributed Computing* (University of Notre Dame Conference Facilities, Notre Dame, IN, USA) *(HPDC '25)*. ACM, Article 12, 13 pages.

[66] Leyi Yan, Linda Wang, Sihang Liu, and Yi Ding. 2025. EnsembleCI: Ensemble Learning for Carbon Intensity Forecasting. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '25)*. ACM, 208–212.

[67] Chaojie Zhang and Andrew A. Chien. 2021. Scheduling Challenges for Variable Capacity Resources. In *Job Scheduling Strategies for Parallel Processing*, Dalibor Klusáček, Walfredo Cirne, and Gonzalo P. Rodrigo (Eds.). Springer International Publishing, Cham, 190–209.

[68] Zhi Zhou, Fangming Liu, Zongpeng Li, and Hai Jin. 2015. When smart grid meets geo-distributed cloud: An auction approach to datacenter demand response. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 2650–2658.

[69] Zhi Zhou, Fangming Liu, Yong Xu, Ruolan Zou, Hong Xu, John C.S. Lui, and Hai Jin. 2013. Carbon-Aware Load Balancing for Geo-distributed Cloud Services. In *2013 IEEE 21st International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems*. 232–241.

A Estimation on Effect of Shifting

A.1 Energy Cost and Carbon Emission

As mentioned in §4.1, we adopt a linear model to estimate power consumption of workloads based on the utilization of each resource type (CPU, GPU, memory) [4, 24, 54, 62]. Specifically, for each job j:

$$P_j^{\text{exec}} = \sum_{c \in C} P_c^{\text{idle}} + (P_c^{\max} - P_c^{\text{idle}}) \times U_{j,c} \tag{7}$$

P_j^{exec} represents the execution power of job j , C is the set of hardware components, $U_{j,c}$ denotes the utilization of component $c \in C$ by job j, and P_c^{idle} and $P_c^{\max}$ are the idle and maximum power consumption of component c, respectively.

Given the power of a job, P_j^{exec}, its energy consumption is calculated based on its duration T_j^{exec} and the data center's PUE:

$$E_j^{\text{exec}} = P_j^{\text{exec}} \times T_j^{\text{exec}} \times \text{PUE} \tag{8}$$

Then, the energy cost of the job running in region d_j is estimated using the average LMP at the region over the job's execution time:

$$\text{Cost}_{j,d_j}^{\text{energy}} = E_j^{\text{exec}} \times \text{LMP}_{d_j} \tag{9}$$

Similarly, the carbon emissions of the job in region d_j are estimated using the average CI at the region over the job's execution time:

$$\text{Carbon}_{j,d_j}^{\text{exec}} = E_j^{\text{exec}} \times \text{CI}_{d_j} \tag{10}$$

A.2 Shifting Overhead

We estimate the overhead of workload shifting in terms of latency, cost, and carbon emissions.

For a job j, the latency of shifting depends on data transmission amount S_j and the number of round trips K_j from the home region s_j to the destination region d_j, bounded by the link bandwidth B_{link} and the propagation delay due to distance between the two regions $\text{RTT}_{s_j \leftrightarrow d_j}$:

$$\text{Latency}_j^{\text{shifting}} = \frac{S_j}{B_{\text{link}}} + \text{RTT}_{s_j \leftrightarrow d_j} \times K_j \tag{11}$$

Cross-region RTT is obtained from prior work repository [59], which was collected from Google Cloud [13]

The cost and carbon impact of shifting consists of two components: data transmission from the home regions s to the destination region d, and the duplicated storage of data in the destination region.

$$\text{Carbon}_j^{\text{shift}} = \text{Carbon}_j^{\text{tran}} + \text{Carbon}_j^{\text{stor}} \tag{12}$$

$$\text{Cost}_j^{\text{shift}} = \text{Cost}_j^{\text{tran}} + \text{Cost}_j^{\text{stor}} \tag{13}$$

We model the energy consumption of data transmission linearly, proportional to the data transmission amount. Since we do not know the exact routing path of the data packets, we use the average CI of the home and destination regions $CI_{s_j \to d_j}^{\text{route}}$ to estimate the carbon emission associated with this transmission:

$$\text{Carbon}_j^{\text{tran}} = S_j \times E_{\text{per GB}}^{\text{tran}} \times CI_{s_j \to d_j}^{\text{route}} \tag{14}$$

where the energy consumption per GB of data transmission is adopted from prior work [24].

Similarly, the energy consumption and carbon emissions of duplicated data storage are estimated based on data volume, the storage device's power P^{stor}, storage duration T_j^{stor}, and the CI at the destination region

$$\text{Carbon}_j^{\text{stor}} = S_j \times P^{\text{stor}} \times T_j^{\text{stor}} \times CI_{d_j} \tag{15}$$

For spatial shifting performed by the user with middleware, the cost of data transmission and storage is simply the listed price, denoted as $R_{\max}^{\text{tran}}$ and $R_{\max}^{\text{stor}}$, respectively. For provider-assisted spatial shifting, the unit cost of data transmission and storage incurred by shifting is interpolated between the lower bound ($R_{\min}^{\text{tran}}$ and $R_{\min}^{\text{stor}}$) and the upper bound ($R_{\max}^{\text{tran}}$ and $R_{\max}^{\text{stor}}$) with a factor $\alpha \in [0, 1]$, as described in § 4.1:

$$R^{\text{tran}} = \begin{cases} R_{\max}^{\text{tran}} & \text{for user} \\ \alpha \times R_{\max}^{\text{tran}} + (1-\alpha) \times R_{\min}^{\text{tran}} & \text{for provider} \end{cases} \tag{16}$$

$$R^{\text{stor}} = \begin{cases} R_{\max}^{\text{stor}} & \text{for user} \\ \alpha \times R_{\max}^{\text{stor}} + (1-\alpha) \times R_{\min}^{\text{stor}} & \text{for provider} \end{cases} \tag{17}$$

Hence, the monetary cost of shifting is estimated as:

$$\text{Cost}_j^{\text{tran}} = S_j \times R^{\text{tran}} \tag{18}$$

$$\text{Cost}_j^{\text{stor}} = S_j \times R^{\text{stor}} \times T_j^{\text{stor}} \tag{19}$$

The data transmission amount, S_j, number of round trips, K_j, and storage duration, T_j^{stor}, depend on data access pattern. Specifically, as described in §4.4, we evaluate three data access patterns:

(1) **Remote I/O**: Data remains in the job's submission region, and all data access operations are performed remotely. Consequently, the total data transferred equals the job's read/write footprints (S_j), the number of network round trips matches the total I/O operations (K_j), and no duplicated storage is required at the destination region ($T_j^{\text{stor}} = 0$).
(2) **Duplication**: Data is copied from the home region to the destination region prior to job execution. This eliminates repetitive network round trip latency ($K_j = 1$), but necessitates storing the data at the destination for the duration of the job's execution ($T_j^{\text{stor}} = T_j^{\text{exec}}$).
(3) **Duplication & Reuse**: This extends the Duplication pattern by caching the data at the destination region for an additional period after the job completes ($T_j^{\text{stor}} = T_j^{\text{exec}} + T^{\text{cache}}$). This enables subsequent repetitive jobs to reuse the cached copy, thereby avoiding redundant data transfers.

Constant values used in our simulations are listed in Table 1.

B Prediction Methods

In addition to simulating workload shifting with an oracle that has perfect future knowledge, we also evaluate a more realistic setting where future grid conditions and workload characteristics are unknown. In particular, we predict grid LMP and CI, as well as the duration and average power consumption of incoming jobs.

B.1 Grid LMP and CI

Although public data sources provide LMP and CI forecasts—such as ISO day-ahead prices [48] and Electricity Maps' three-day forecasts [38]—0.96% of jobs run longer than one day, and 0.08% exceed three days. We made prediction for CI and LMP using LightGBM [29], with weather data as covariates. Compared to simulations with oracle grid information, using these predicted CI and

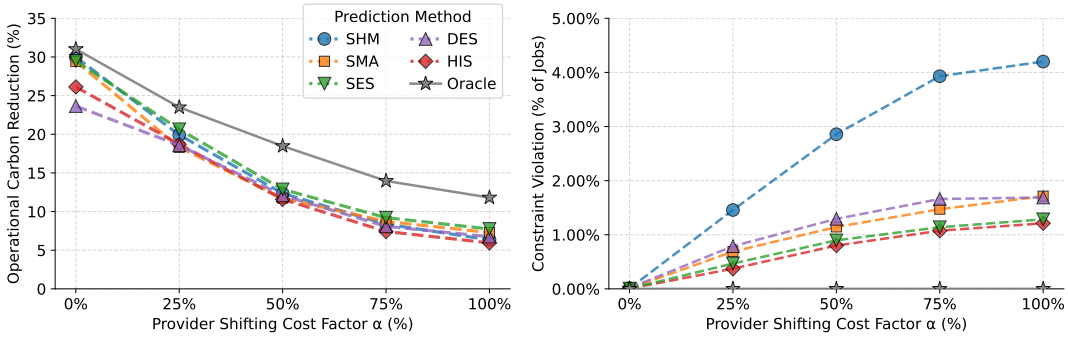

Figure 8: Impact of different prediction methods on operational carbon reduction (left) and constraint violation rates (right) across varying provider shifting cost factors.

Table 1: Constant values used in our simulations.

Item	Value
CPU Price [21]	$0.03465/vCPU-hour
Memory Price [21]	$0.003938/GB-hour
GPU Price - T4 [23]	$0.35/hour
GPU Price - P100 [23]	$1.46/hour
GPU Price - V100 [23]	$2.48/hour
Data Egress Price [22]	$0.02/GB
Data Storage Price [18]	$1.7e-5 GB/hour
Provider Network Cost [10]	$0.05 Mbps/month
Provider Storage Cost [53]	$85.62 e3/EB/10-year
CPU Power (64 cores) [1]	360 W
GPU Power (96 cores) [1]	400 W
Mem Power [11]	0.838 W/GB
GPU Power - T4 [43]	70 W
GPU Power - P100 [42]	250 W
GPU Power - V100 [44]	250 W
Storage Power [53]	733.855 kW/EB
Network Energy [24]	0.005 kW/GB
Power Usage Effectiveness (PUE) [19]	1.1
Link Bandwidth [51]	5 Gbps

LMP values yields similar operational carbon savings while maintaining a constraint violation rate less than 1% (i.e., incorrect job-shifting decisions that exceed the budget or latency constraint due to prediction errors).

B.2 Job Duration and Power Consumption

For each incoming job, we assume the either the user specifies its data requirements using the interface described in §3.1, or it is known (e.g., using static analysis). We then predict the job's duration and power consumption based on historical execution data. The workload trace dataset groups repetitive job submissions with similar meta-information, such as entry scripts, command-line parameters, and data sources or sinks [64]. We aggregate the historical duration and power consumption for each of these groups. For jobs outside previously observed groups, we fall back to per-user historical statistics. If the user is also unknown, we use global statistics across all jobs.

Because grid information is collected and predicted hourly while jobs arrive at a much higher frequency, we use lightweight statistical forecasting methods instead of computationally intensive training-based approaches to minimize overhead. We evaluate five statistical predictors: Simple Historical Mean (SHM), Simple Moving Average (SMA), Simple Exponential Smoothing (SES), Double Exponential Smoothing (DES), and Histogram (HIS). We compare their carbon savings and constraint violation rates, shown in Fig. 8. Among these methods, SES achieves the best overall performance, with the best carbon reduction and a maximum constraint violation rate of 1.2%. We therefore use SES to forecast job duration and power consumption throughout the paper.