CDN-Shifter: Leveraging Spatial Workload Shifting to Decarbonize Content Delivery Networks

Jorge Murillo
University of Massachusetts Amherst
Amherst, MA, USA
jrmurillo@umass.edu

Walid A. Hanafy
University of Massachusetts Amherst
Amherst, MA, USA
whanafy@cs.umass.edu

David Irwin
University of Massachusetts Amherst
Amherst, MA, USA
irwin@ecs.umass.edu

Ramesh Sitaraman
University of Massachusetts Amherst
Amherst, MA, USA
ramesh@cs.umass.edu

Prashant Shenoy
University of Massachusetts Amherst
Amherst, MA, USA
shenoy@cs.umass.edu

ABSTRACT

Content Delivery Networks (CDNs) are Internet-scale systems that deliver streaming and web content to users from many geographically distributed edge data centers. Since large CDNs can comprise hundreds of thousands of servers deployed in thousands of global data centers, they can consume a large amount of energy for their operations and thus are responsible for large amounts of Green House Gas (GHG) emissions. As these networks scale to cope with increased demand for bandwidth-intensive content, their emissions are expected to rise further, making sustainable design and operation an important goal for the future. Since different geographic regions vary in the carbon intensity and cost of their electricity supply, in this paper, we consider spatial shifting as a key technique to jointly optimize the carbon emissions and energy costs of a CDN. We present two forms of shifting: spatial load shifting, which operates within the time scale of minutes, and VM capacity shifting, which operates at a coarse time scale of days or weeks. The proposed techniques jointly reduce carbon and electricity costs while considering the performance impact of increased request latency from such optimizations. Using real-world traces from a large CDN and carbon intensity and energy prices data from electric grids in different regions, we show that increasing the latency by 60ms can reduce carbon emissions by up to 35.5%, 78.6%, and 61.7% across the US, Europe, and worldwide, respectively. In addition, we show that capacity shifting can increase carbon savings by up to 61.2%. Finally, we analyze the benefits of spatial shifting and show that it increases carbon savings from added solar energy by 68% and 130% in the US and Europe, respectively.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SoCC '24, November 20–22, 2024, Redmond, WA, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1286-9/24/11
https://doi.org/10.1145/3698038.3698516

CCS CONCEPTS

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

KEYWORDS

Load shifting, Decarbonization, Internet-scale Content Delivery

ACM Reference Format:
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 *ACM Symposium on Cloud Computing (SoCC '24), November 20–22, 2024, Redmond, WA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3698038.3698516

1 INTRODUCTION

Modern Internet services have long relied on globally-deployed distributed systems to serve their users. With the emergence of global-scale commercial cloud providers and ever-increasing demand from application providers, the demand for cloud computing has grown significantly over the past two decades [4]. At the same time, the rise of latency and bandwidth-sensitive Internet services has given rise to edge services that use computing and storage resources at the "edge" of the Internet to deploy latency-sensitive and bandwidth-sensitive content. Internet content delivery is an

early example of an edge service that delivers various types of content, such as video and web content, from a distributed network of edge servers. Analogous to hyperscaler cloud platforms with a global footprint, large content delivery networks (CDNs) use edge platforms comprising hundreds of thousands of servers distributed across the globe. For example, the Akamai CDN comprises approximately 350,000 servers in 134 countries and over 1,300 networks worldwide [3]. A vast amount of Internet traffic runs through CDNs, with the Akamai CDN serving hundreds of billions of web requests per day in the year 2022 [47] and 56.7% of the 10000 most popular sites use a CDN [25, 38].

Since data centers account for nearly 3% of global carbon emissions, with this figure set to rise significantly in the coming decade, the sustainability of cloud platforms has received significant attention in recent years. Major cloud providers have announced aggressive goals to become zero carbon by 2030 [23, 42, 43]. Also, researchers have proposed numerous techniques, such as incorporating renewable energy [31, 33] and making operations carbon-aware to reduce the carbon footprint of cloud platforms. Since global CDNs rival large-scale cloud platforms in their size and scale, the sustainability of CDNs is also important, but has seen less attention. While some early work has emphasized reducing the energy footprint of CDNs [18, 31, 33, 35], these optimizations were focused on reducing the energy and operational cost of these platforms and did not emphasize the problem of sustainability and the reduction of the carbon footprint of CDNs. It is worth noting that optimizing a computing system's carbon footprint entails more than reducing its energy use. Doing so involves increasing the use of low-carbon, or carbon-free, energy which in turn reduces the system's carbon emissions [5]. However, carbon optimization techniques developed for cloud platforms do not directly apply to edge platforms such as Internet-scale CDNs for the following reasons.

First, many recent approaches for reducing operational carbon emissions of the cloud target batch workloads, such as machine learning training and scientific applications [16, 44, 50]. Since the carbon intensity of electricity supply varies over time, such approaches initially focused on exploiting these variations through *temporal* workload shifting, where the batch workload demand is shifted from high to low carbon periods to reduce carbon consumption [19, 20, 50]. Temporal workload shifting is unsuitable for CDNs since CDN workloads are latency-sensitive interactive services that must be serviced in real-time. Second, some recent approaches have focused on the spatial shifting of batch workloads, such as machine learning training to greener cloud regions [9, 46]. While shifting workloads to a distant cloud region to reduce carbon emissions is suitable for many cloud batch workloads with less stringent completion time requirements, these methods do not directly apply to CDN workloads since they have strict performance requirements due to their latency-sensitive nature. Specifically, any carbon-aware CDN optimization must be cognizant of any potential response time degradation from sending interactive requests to CDN edge server locations far from the end user. Third, since CDNs are geographically distributed on a global scale, they incur significantly different operating costs, in terms of electricity costs, across regions. Hence, it is imperative that any carbon-aware optimization not cause an inadvertent increase in electricity costs, requiring techniques to consider both carbon and energy costs.

The carbon-aware CDN optimizations presented in the paper are motivated by two insights. First, the carbon intensity of electricity exhibits spatial variations that can be exploited by a CDN. This is because energy generation at different locations uses different mixes of generation sources such as solar, hydro, wind, and coal, yielding spatial differences in the carbon intensity [10, 49]. Importantly, these variations should be exploited while also considering spatial differences in electricity prices. Second, CDNs have built-in spatial redundancy to serve the same content to users from several edge locations. A CDN's global load balancer uses this spatial flexibility to serve content to users from the closest suitable edge location. These two insights motivate using spatial load-shifting methods to reduce a CDN's carbon footprint. For example, if two nearby edge data centers cache the same content, but one has a lower carbon electricity supply than the other, then serving users requesting content from the greener location will incur lower carbon consumption at the possible expense of a slightly higher user latency. So long as the latency increase is small, adding such spatial carbon awareness into the CDNs load balancer remains promising.

In this paper, we present new carbon-aware spatial shifting approaches to enhance the sustainability of a global CDN platform. Our shifting approach utilizes the knowledge of the demand, carbon intensity, and energy costs to shift CDN workloads across edge data centers. Specifically, we design optimization-based carbon- and cost-aware approaches to shift both load and capacity to greener regions while minimizing the latency impact on end users. In doing so, our work asks the following questions: *To what extent can the carbon emissions of an Internet-scale CDN be reduced with carbon-aware spatial load shifting, and what is the possible latency increase? How can such load shifting jointly optimize carbon emissions and operational electricity costs? How can a CDN redistribute its capacity to greener regions to enhance the efficacy of spatial load shifting? Finally, what benefits can the use of local renewable energy bring, and how much load shifting extend their benefits?*

To address the above questions, our paper presents CDN-Shifter that leverages different types of spatial shifting to

decarbonize large CDNs. In designing and evaluating CDN-Shifter, we make the following contributions:

- We formulate the carbon-aware load shifting for CDNs as an optimization problem to jointly optimize carbon emissions and electricity costs. We design a carbon-aware load balancer that leverages the optimization problem to reduce carbon footprint and energy costs by spatially shifting workloads.
- While load shifting operates at a time scale of minutes, we present a second form of shifting—capacity shifting–that operates at a coarser time scale of days or weeks to move CDN capacity to green regions with constrained capacity. Such carbon-aware capacity shifting can improve the efficacy of load-shifting methods.
- We use the real-world traces from Akamai CDN comprising more than 2,600 geographically distributed edge sites and carbon and energy costs across the globe to analyze the benefits and limitations of carbon- and cost-aware shifting as well as methods to further increase such savings.
- Our evaluation results demonstrate that for a 60ms latency increase, carbon-aware spatial load balancing achieves up to 35.5%, 78.6%, and 61.7% carbon savings across the US, Europe, and worldwide, respectively. In addition, we demonstrate how capacity shifting increases carbon savings by up to 61.2%. Finally, we analyze the benefits of combining spatial shifting with adding renewables and how it can increase carbon savings from added solar energy by 68% and 130% across the US and Europe, respectively.

2 BACKGROUND

In this section, we provide a background on content delivery networks, the carbon intensity of the grid's electricity, electricity prices, and temporal and spatial shifting.

2.1 Content Delivery Networks (CDNs)

Content delivery networks are large-scale edge systems that use a network of edge data centers to deliver content such as web pages and video and audio streams. Commercial CDNs, such as the Akamai CDN, are characterized by having a *global deployment* consisting of thousands of edge data centers and hundreds of thousands of servers deployed in those locations, and a massive *replication* of services across these data centers. [34]. To ensure high levels of coverage, CDNs utilize a mixture of dedicated data centers, co-located data centers, and virtual clusters in public and private clouds [15]. For instance, CDN traces from Akamai, a major CDN provider with more than 113k servers worldwide, show that 50% of the edge sites have less than ten servers, indicating the breadth of their deployment options.

CDNs use a two-level approach to service incoming requests. First, each incoming request is mapped by a *global* load-balancing algorithm to an edge data center location. Next, a *local* load balancer maps the request to a specific server within that data center, which then services that request using cached content. In many cases, the global load balancer is embedded into the Domain Name Service (DNS)'s lookup process, where the DNS maps the request to the nearest or the most suitable edge data center, thereby performing global load balancing. A key consideration is the performance (e.g., latency) seen by interactive requests, which usually means that requests should be serviced from nearby data centers to provide low-latency service.

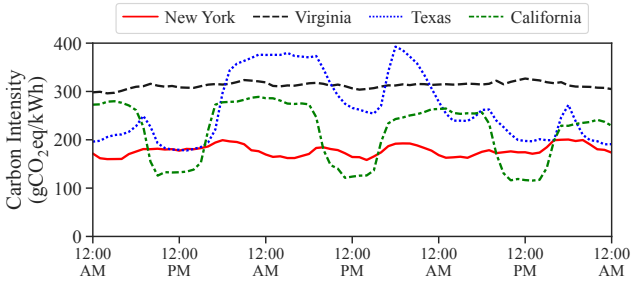


Figure 1: Carbon Intensity for four Electricity Grids in the US between February 10-12.

Similar to large cloud platforms that consume a significant amount of energy to power and cool their servers, the large number of edge data centers and servers comprising a CDN can consume significant amounts of energy. Thus, similar to cloud providers, CDNs have considered approaches to minimize the carbon footprint and cost by utilizing spatial flexibility, as content delivery applications are mostly stateless. The global load balancer can exploit this flexibility to redirect requests to edge locations with the greenest/cheapest energy supply rather than to the nearest edge location available. In doing so, CDNs have the potential to reduce both their operations' carbon emissions and costs.

Because both the output of renewable energy sources varies through time (mainly due to weather conditions), the electric grid will choose to activate other power sources to match demand, which also varies over time.

2.2 Carbon Intensity of Electricity

Carbon Intensity is a metric used to describe the amount of greenhouse gasses an electric grid releases to the atmosphere per unit of energy used/generated, generally described in units of grams of CO2 equivalent per kilowatt-hour (gCO2eq/kWh). This metric is determined by the mixture of power generation sources that an electric grid uses at a specific time. These sources can include coal, gas, and nuclear plants, which release substantial amounts of CO2 into the atmosphere, as well as renewable sources such as hydro

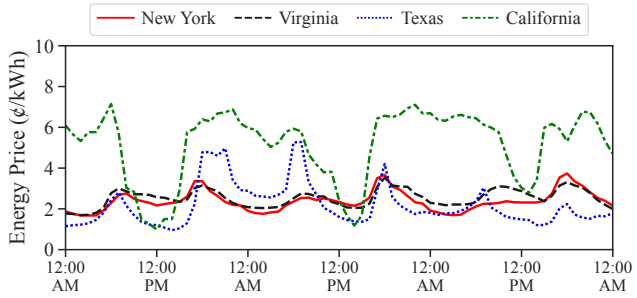


Figure 2: Energy Prices for four Electricity Grids in the US between February 10-12.

plants, wind turbines, and solar panels, which are essentially carbon-free. The electric grid will choose to activate different power sources at different points in time because of two factors: the power output of renewable sources is intermittent (as they are subject to weather conditions), and user demand for electricity varies with time. This results in the carbon intensity of the electricity from a single grid to fluctuate over time. Electricity grids in different geographical regions deploy different kinds and numbers of power plants and have different weather patterns, which results in different regions having entirely different carbon intensity profiles. Figure 1 depicts carbon intensity measured in g·CO_2eq/kWh in four electricity grids in the US. As shown, carbon intensity highly differs across locations and times as the energy sources change. For example, California's carbon intensity follows the notable diurnal duck curve [6] due to its high dependency on solar energy. In contrast, New York and Virginia highly depend on gas and, hence, have more stable carbon intensity. In addition, the figure shows that carbon intensity in California differs temporally by up to 2.3× and spatially by up to 2.5×.

2.3 Electricity Prices

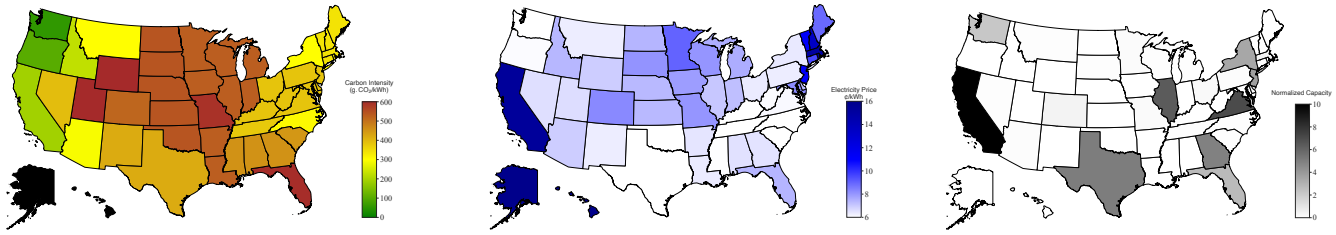
Matching energy supply and demand throughout the day ensures the stability of the electricity grid. To do so, electricity grids often match supply and demand by running an energy market where consumers can purchase energy in advance (e.g., Day-ahead markets) or in real-time markets [22, 41]. Electricity grid and providers utilize multiple pricing models. For example, large-scale (e.g., utility companies or data centers) consumers often obtain PPAs and buy energy from day-ahead markets to ensure stable prices. In addition, they often utilize real-time, also called spot energy market, to compensate for the deficiency or sell excess energy. In contrast, small consumers often depend on their local utility company, which sets a fixed or a time-of-use electricity price. Temporal and spatial variations in supply and demand result in energy price differences across states and throughout the day. For instance, the high availability of almost cost-free renewables often decreases the energy price. Similarly, when demand is low, e.g., during the night, electricity prices significantly drop. Figure 2 shows an example of day-ahead energy prices in four electricity grids within the US. As shown, prices vary differently across locations and times. For instance, most grids exhibit lower prices at night as demand is usually lower. In contrast, California, which highly relies on solar energy, depicts an opposite behavior, where energy prices are often higher at night due to the lack of solar energy production. CDNs have exploited spatial differences in electricity prices across regions to reduce their operational costs by moving loads to locations with lower electricity prices [33, 37]. However, such optimizations did not consider carbon costs, which is the focus of our work.

2.4 Carbon-aware Load Shifting

To exploit variations in energy's carbon intensity, cloud computing platforms have employed *temporal* and *spatial* load-shifting techniques to reduce total carbon emissions. Temporal workload shifting exploits the time-varying nature of a grid's carbon intensity by delaying workloads to times when carbon intensity is lower. Temporal workload shifting is generally employed with batch jobs, as they have more permissible completion times and are generally less latency-sensitive than interactive workloads [19–21, 46, 50]. Of course, due to the interactive nature of CDN requests, temporal shifting is not well suited for CDN workloads.

In contrast to temporal shifting, spatial shifting involves taking advantage of geographical variations in carbon intensity. For instance, a spatial workload shifting strategy for batch workloads could involve relocating tasks to remote cloud regions with access to green energy sources [46]. Moreover, spatial shifting can also be applied to interactive workloads like web services and video streaming by moving them to regions with greener energy sources [9, 30, 45]. However, shifting workloads to farther away locations often introduces latency overheads.

Our work considers spatial workload shifting in the context of a CDN's geographic load shifting. Specifically, we consider the extent to which the load balancer can shift work to data centers with greener and cheaper energy while also being aware of performance considerations. Such performance considerations can limit the geographic radius where requests from a certain region can be moved. In addition, previous research on carbon-aware load shifting for cloud workloads has only considered carbon optimizations. In the CDN case, additional considerations arise. For example, moving the workload load to lower carbon-intensity regions might increase total operational costs, as regions with low carbon



(a) Carbon Intensity (g·CO_2eq/kWh) (b) Electricity Price (¢/kWh) (c) CDN Capacity

Figure 3: Carbon Intensity (a), Electricity Prices (b), and CDN Capacity (c) illustrate the diversity across US states.

intensity may have higher energy costs. This raises a three-way trade-off between carbon emission, electricity cost, and request latency that has not been addressed previously.

To our knowledge, there is only one work that has considered such a three-way trade-off for CDNs [12] through a data placement and load-shifting heuristic. However, their work assumed fixed carbon intensity and costs per region, whereas carbon intensity varies continuously over time in practice. Further, this work did not consider the impact of local renewables and other forms of shifting, such as capacity shifting, to reduce emissions and costs further. In Section 3, we further motivate and differentiate our work by considering the conflict between the availability of green and cheap electricity as well as opportunities and challenges of reducing global carbon emissions using real-world, carbon, cost, and CDN traces.

2.5 Carbon-aware Capacity Shifting

Today's CDNs use diverse types of infrastructure, the majority of which is distributed among co-located data centers and virtual clusters in public and private clouds. For example, 5G networks and cloud-based CDNs often rely on fully virtualized resources [2, 15]. In scenarios where CDN servers are virtualized, additional carbon-aware optimizations become feasible. For example, load shifting to green regions will naturally increase the utilization in those regions. At high utilization levels, additional load shifting becomes challenging due to capacity constraints at those edge locations. However, in the case of virtualized architectures, it is possible to quickly provision additional virtual machines (and storage) at edge locations to temporarily increase the capacity of edge locations in green regions. Doing so can enable additional load shifting to reduce carbon emissions in a cost and latency-aware manner. Our work considers such capacity-shifting optimization to augment our load-shifting approach. We note that other efforts have considered the notion of capacity in cloud platforms [2, 13, 24, 39], but these efforts did not consider limitations on moving capacities and their size.

3 DESIGNING SUSTAINABLE CONTENT DELIVERY NETWORKS: MOTIVATION

The size of CDNs highlights the magnitude of their energy consumption and emissions. In this section, we motivate the opportunities and challenges in cost- and carbon-aware spatial shifting. We do so using real-world carbon, cost, and CDN traces. See Section 5.1 for more details about the traces and data preprocessing.

Figures 3a, and 3b presents the spatial diversity of grid-supplied energy in the US in June 2021. Figure 3a shows the variations in average carbon intensity across the US.[1] As shown, carbon intensity exhibits a high level of heterogeneity where the ratio between the state with the greenest energy (Washington) and brownest energy (Utah) is 10.6×. In addition, the coefficient of variation (σ/μ) in carbon intensity across the US is 0.3. Figure 3b shows energy retail prices across the US during June 2021 for industrial consumers. Similarly, energy prices exhibit a high level of variation, where the ratio between Texas and California states is 3× with a coefficient of variation across all states of 0.46. The diversity in both carbon intensity and costs motivates the opportunities for spatial shifting. For example, shifting to greener regions can drastically reduce carbon emissions.

Although individually, energy's carbon intensity and prices make the case for spatial shifting to decrease total operational carbon emissions or costs. Naively shifting workloads based on carbon intensity or energy prices alone may lead to large penalties. For example, if we consider requests from California and Utah, a carbon-aware policy will prioritize CDNs in California, as energy is greener but will increase total operational costs. In contrast, a cost-aware policy will prioritize CDNs in Utah, as energy is much cheaper, which will increase total carbon emissions. Thus, load-shifting decisions must consider the trade-off between carbon emissions and costs. We detail the breadth of such conflict in Section 5.4.3.

Moreover, typical CDN load balancers often focus on minimizing latency by scheduling requests to the nearest CDN.

[1] Carbon Intensity for Alaska and Hawaii were not available.

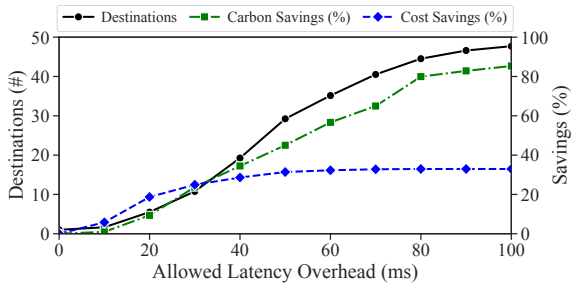


Figure 4: Possible destinations and savings at different latency limits in June 2023.

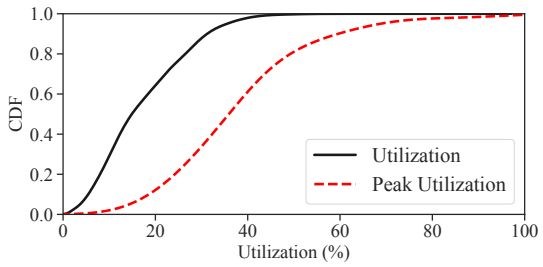


Figure 5: Peak and average utilization in CDNs across the US.

In contrast, a carbon-aware load balancer will consider trading latency increases for carbon savings. Figure 4 shows the relation between latency overheads, flexibility (number of allowed destinations), and average savings. As shown, higher latency overheads correlate with increased destinations (shifting opportunities) and increased savings. For example, a 50ms latency overhead may enable 45% and 35% carbon emissions and cost savings, respectively. While the allowed latency overheads are application-specific, we envision CDNs being more tolerant to latency overheads than latency-critical applications such as AR and VR. Nonetheless, load-shifting techniques must consider CDNs performance requirements and the trade-offs between savings and performance.

While Figure 4 shows an ideal scenario where destination locations have no capacity constraints. CDNs, like other data centers, have capacity constraints, and their capacity and utilization levels dictate the scheduling flexibility. Figure 5 depicts the utilization distribution among all edge sites. The results show that CDNs are typically underutilized, with a mean utilization of 17%. The results also show that CDNs peak utilization barely reaches the full capacity of the CDN, with an average of 37%. These low utilization rates depict the potential of spatial shifting, where the load is shifted to decrease operational costs.

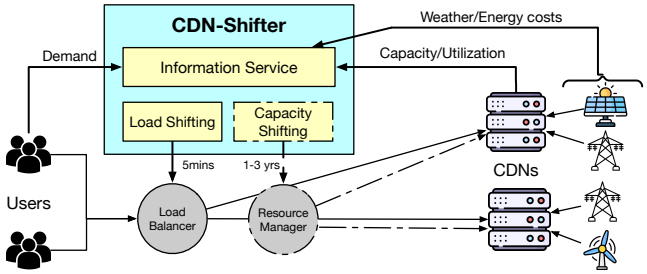


Figure 6: CDN-Shifter Design.

Utilization rate only conveys part of the limitations of spatial shifting, where actual available capacity is also crucial for shifting. Figure 3(c) shows the capacity variations across the US states. As shown, the figure shows a high degree of variation across states. For example, California, Texas, and Virginia hold 34% of the total capacity in the US. This concentrated capacity leads to missing many opportunities to shift load. For example, compared to Virginia, North Carolina has greener and cheaper energy but negligible capacity. Similarly, Washington State has cheaper and greener energy than California but a much smaller capacity. The trends in CDN distributions and energy profiles across US states motivate possible ways to reduce operational carbon emissions and costs by redistributing the capacity (i.e., capacity shifting) in the most impactful way.

In addition to opportunities for shifting loads and capacity, renewable energy plays a crucial role in reducing operational emissions. For instance, CDN operators can introduce renewable energy-based solutions in areas with brown energy sources. However, as demonstrated in previous studies [1, 7], the effectiveness of renewable energy diminishes significantly. This is due to the intermittent nature of most renewable energy sources. For example, integrating solar energy can meet the demand only during daytime hours. Therefore, operators must combine their efforts by adding renewables with load-shifting techniques to reduce their total carbon emissions. The key takeaways are:

(1) Diversity across geographical locations offers the potential to decrease operational carbon emissions and costs.
(2) Load shifting decisions must consider the three-way trade-off between latency, carbon emissions, and cost while adhering to capacity constraints.
(3) The unused capacity across edge data centers can be redistributed to further decrease carbon emissions.
(4) Load shifting is essential to increase the efficacy of added renewables.

4 CDN-SHIFTER DESIGN AND POLICIES

In this section, we will outline the architecture of CDN-Shifter. Next, we introduce an optimization approach for carbon- and cost-aware spatial workload shifting, where we move the CDN load from green to brown regions. Finally, we present a capacity-shifting approach that relocates the provisioned capacity across data centers.

4.1 System Architecture

Figure 6 shows the architecture of CDN-Shifter where users are grouped based on location. CDN servers are typically hosted in dedicated, co-located, and public data centers, where content is replicated according to the expected demand. The CDN global load balancer maps requests to the nearest CDN where content is available. CDN-Shifter extends the global load balancer by integrating data from different sources in order to enhance its placement decisions but does not interfere with the load balancer operations. CDN-Shifter's scheduling decisions are integrated into the global load balancer, where users are mapped to edge data centers. In addition, CDNs often rely on resource managers, which are responsible for operating the underlying infrastructure. CDN-Shifter extends the resource manager by implementing the resource provisioning decisions. CDN-Shifter implements three key components:

Information Service: CDN data centers often use a variety of energy sources, including the local electricity grid and renewable sources such as solar farms. Our system, CDN-Shifter, incorporates an Information Service that monitors carbon emissions and energy prices. The Information Service relies on weather forecasts to assess the availability of renewable energy and computes the expected excess demand, which will be satisfied using the local grid. Additionally, it tracks the real-time carbon emissions and energy prices from the electricity grid (using services such as ElectricityMaps [10] and real-time energy markets). Lastly, our Information Service forecasts the demand, usage, and capacity of CDNs to identify opportunities for load shifting.

Load-Shifting: The load-shifting policy operates at a fast time scale of minutes. It determines how load, based on users' locations, is mapped to CDN locations at fine-grained granularity (e.g., five minutes). One way to implement such a policy is to iterate over locations where the content is available and the expected latency is lower than the SLO (Service-Level Objective) threshold and forward requests to the CDN with the cheapest energy. However, such a greedy approach ignores global reductions. In Section 4.2, we formalize the load-shifting optimization problem to minimize total operational carbon emissions and costs while respecting the latency and capacity constraints.

Table 1: *Load Shifting Parameters and Decision Variables.*

Notation	Description
N	$N = \{0, 1, ..., n\}$ is a Set of CDN data centers.
d_{ijt}	Latency between data centers i and j at time t.
N_{it}^{δ}	$N_{it}^{\delta} \subseteq N$ Set of data centers with latency $d_{ijt} \leq \delta$, including i.
l_{it}	Incoming Load to data center i at time t.
c_{it}	Resource capacity of data center i at time t.
PUE_i	Power Usage Efficiency at data center i.
E_i	Energy consumption per unit load at data center i.
CI_{it}	Data center i energy's carbon intensity at time t.
P_{it}	Data center i energy's price at time t.
α	Carbon-Cost balance factor.
λ	Latency overhead factor.
L_{ijt}	Load shifted from data cente i to data cente j.

where $i, j \in N$

Capacity-Shifting: The capacity shifting policy operates at slower time scales of days or weeks. It determines how capacity should be opportunistically provisioned at green CDN locations to maximize the benefits of the above load shifting policy. In Section 4.3, we formalize the capacity-shifting optimization problem to minimize total operational carbon emissions and costs while respecting capacity and demand constraints.

4.2 Carbon-aware Load Shifting

Consider a CDN network with N edge data cente with their geographic locations. In each time slot (e.g., five minutes), each CDN location receives some load and has a finite resource capacity. We consider time-varying resource capacity to include server failures and upgrades. To optimize the total operational carbon emissions, the CDN's global load balancer must consider the energy's carbon intensity and price to determine how much load to move from one edge data cente to another to minimize the total operational emissions while serving the total workload. We model this problem as a linear optimization problem that needs to be solved by the global load balancer in each time slot. Table 1 describes the used input parameters and decision variables.

The objective can be written as a minimization problem where CDN-Shifter seeks to minimize the total operational emissions and costs of the entire system at each time step t.

$$\text{Minimize} \sum_{i=1}^{N} \sum_{j \in N_{it}^{\delta}} L_{ijt} \times \boldsymbol{\mathcal{C}}_i \tag{1}$$

$$\boldsymbol{\mathcal{C}}_i = E_i \times PUE_i \times (\alpha\, CI_{it} + (1 - \alpha)\, P_{it})$$

s.t.

$$\sum_{i \in N_{jt}^{\delta}} L_{ijt} \leq c_{jt}, \quad \forall j, t \tag{2}$$

$$\sum_{j \in N_{it}^{\delta}} L_{ijt} = l_{it}, \quad \forall i, t \tag{3}$$

$$L_{ijt} \geq 0, \quad \forall i, j, t \tag{4}$$

where $\mathcal{C}_i$ is the operational carbon emissions and monetary costs at data center i. E_i, PUE_i denotes the servers' and data center' energy efficiency. Note that although our model assumes energy consumption is linear to the load across data centers, incorporating energy importionality is straightforward. $\alpha \in [0, 1]$ is the carbon-cost balance factor, where $\alpha = 1$, yields a carbon-aware load shifting policy, while $\alpha = 0$ yields a cost-aware load shifting policy. Observe that any performance constraint in terms of the maximum latency that can be used to shift workload and limit the latency increase is captured using N_{it}^{δ}, which is the feasible set of nearby locations for each edge data center i subject to a specific latency constraint δ, where we assume that d_{ijt} is time variable to account for daily latency variations. Equation 2 guarantees that the incoming load to a data center does not exceed its resource capacity. Equation 3 is the load conservation constraint, which ensures that the outgoing load from a data center should equal the initial load from the data center. This includes the load from a data center to itself (i.e., the load that stays at a data center). Finally, Equation 4 states that the load transfers should be non-negative.

While the above formulation yields feasible solutions, it suffers from one problem: there are multiple ways to set the load transfers L_{ijt} to get the same optimal objective, each of which will result in a different value for the overall latency increase in the system. Although this problem can be solved using a two-step lexico-graphic optimization. We chose to directly augment the objective function with the latency overhead using a latency overhead penalty factor λ, as follows:

$$\text{Minimize} \sum_{i=1}^{N} \Big(\underbrace{\sum_{j \in N_i^{\delta}} L_{ijt} \times \mathcal{C}_i}_{\text{Operational Costs}} + \lambda \underbrace{\sum_{j \in N_i^{\delta}} L_{ijt} \times d_{ij}}_{\text{Latency Penalty}} \Big) \tag{5}$$

Finally, we note that CDN-Shifter solves this optimization problem every time step, where energy's carbon intensity and price, as well as total load and capacity, can be accurately estimated.

4.3 Carbon-aware Capacity Shifting

CDNs are composed of physical and virtual resources, and load shifting, as explained earlier, utilizes the current set of edge clusters of a CDN network to jointly optimize its carbon and electricity costs. However, carbon savings are limited by latency constraints as well as capacity constraints at the locations with the greenest energy sources. In the presence of load shifting, some edge locations (e.g., with high carbon or energy costs) may experience low utilization, while green edge locations may become highly utilized, leaving them unable to accept additional load from other locations. To address such issues, our work employs capacity shifting where virtual machine capacity can be intelligently provisioned in greener, highly utilized regions and deprovisioned from under-utilized regions to lower overall costs. Such dynamic provisioning has long been studied for web-based cloud applications, but those techniques are designed for a single location, while our approach performs cross-site provisioning across CDN edge locations. We formulate this problem as an optimization problem considering future carbon intensity and prices, users' demand, and infrastructure (e.g., buildings and power) capacity. Table 2 describes the used input parameters and decision variables. Note that we implement the capacity migrations based on expected demand and operational costs. The objective can be written as a minimization problem where CDN-Shifter optimizes the expected operational costs for an upcoming time horizon.

Table 2: *Capacity Shifting Parameters and Decision Variables.*

Notation	Description
N	$N = \{0, 1, ..., n\}$ is a Set of CDN data centers.
d_{ij}	Latency between data centers i and j
N_i^{δ}	$N_i^{\delta} \subseteq N$ Set of data centers with latency $d_{ij} \leq \delta$, including i.
l_i	Average load at data center i.
l_i^{max}	Peak load at data center i.
c_i	Resource capacity of data center i.
ψ_i	$\psi_i \geq 1$ is the capacity expansion factor of data center i.
PUE_i	Power Usage Efficiency at data center i.
E_i	Energy consumption per unit load at data center i.
CI_i	Average carbon intensity at data center i.
P_i	Average energy price at data center i.
α	Carbon-Cost balance factor.
λ	Latency overhead factor.
γ	Capacity shifting factor.
L_{ij}	Load Shifted from data center i to data center j.
C_{ij}	Moved Capacity from data center i to data center j.

where $i, j \in N$

$$\text{Minimize} \sum_{i=1}^{N} \sum_{j \in N_i^{\delta}} L_{ij} \times \mathcal{C}_i \tag{6}$$

$$\mathcal{C}_i = E_i \times PUE_i \times (\alpha\, CI_i + (1 - \alpha)\, P_i)$$

s.t.

$$\sum_{j \in N_i^\delta} L_{ji} \leq \sum_{j \in N} C_{ji}, \quad \forall i \tag{7}$$

$$l_i^{max} \leq \sum_{j \in N} C_{ji}, \quad \forall i \tag{8}$$

$$\sum_{j \in N} L_{ij} = l_i, \quad \forall i \tag{9}$$

$$\sum_{j \in N} C_{ij} = c_i, \quad \forall i \tag{10}$$

$$\sum_{j \in N} C_{ji} \leq \psi_i \, c_i, \quad \forall i \tag{11}$$

$$L_{ij}, C_{ij} \geq 0 \tag{12}$$

Similar to the previous section, $\boldsymbol{C}_i$ is the operational costs to include carbon emissions and monetary costs at data center i, E_i, PUE_i are server's and data center's energy efficiency. $\alpha \in [0, 1]$ is the carbon-cost balance factor, and N_i^δ is used to enforce latency constraints δ. Equation 7 motivates capacity migration to serve the newly shifted load. Equation 8 ensures that each data center can still serve its peak demand without load shifting. Equation 9 is the load conservation constraint, where all load must be met. Equation 10 enforces capacity conservation, where total resources are maintained. Equation 11 guarantees that the new data center's capacity does not surpass the power and infrastructure capacity limits, where $\psi_i \geq 1$ is the available headroom for expansions at data center i. Equation 12 states that the load and capacity transfers should be non-negative. Finally, to ensure latency minimization and avoid unnecessary capacity shifts, we augment Equation 6 with the latency overhead using a latency overhead penalty factor λ and shifting overhead using a capacity shifting overhead penalty factor γ as follows[2]:

$$\text{Minimize} \sum_{i=1}^{N} \Big(\underbrace{\sum_{j \in N_i^\delta} L_{ij} \times \boldsymbol{C}_i}_{\text{Operational Costs}} + \lambda \underbrace{\sum_{j \in N_i^\delta} L_{ij} \times d_{ij}}_{\text{Latency Penalty}} + \gamma \underbrace{\sum_{j \in N, j \neq i} C_{ij}}_{\text{Shift Penalty}} \Big). \tag{13}$$

5 EVALUATION

In this section, we evaluate the potential of spatial load shifting, capacity shifting, and adding solar energy in content delivery networks (CDNs). In doing so, we answer the following questions:

(1) What is the potential for reducing operational carbon emissions and costs through spatial load shifting?
(2) What is the breadth of the trade-off between carbon reductions and operational costs?
(3) How can capacity shifting further reduce operational emissions or costs?
(4) How does load shifting amplify the benefits of added renewables?

Next, we outline our real-world datasets, experimental setup, and evaluation metrics.

[2] γ can also be used as a shifting cost across data centers.

Table 3: *The location, number of hosts, and the total number of sites within that location.*

Zone	Hosts (#)	Sites (#)	Carbon Intensity (g·CO_2eq/kWh)	Energy Price (¢/kWh)
North America	64.3k	1327	438.90	8.46
Central America	0.2k	29	245.23	27.36
South America	2k	121	169.03	47.93
Europe	31.9k	585	304.24	10.22
Asia	12.7k	486	521.50	11.76
Oceania	2k	114	447.86	16.75
Africa	0.3k	29	713.12	21.08
Worldwide	113k	2691		

Table 4: *Parameters for generating solar energy traces using PVWatts tool [8].*

Parameter	Value	Unit
DC System Size	1	kW
Azimuth	180 (the northern hemisphere)	deg
	0 (the southern hemisphere)	deg
Tilt	| latitude |	deg
System Losses	14.08	%
Module Type	Standard	
Array Type	Fixed (open rack)	

5.1 Real-world Datasets

Our evaluation setup uses real-world CDN, carbon intensity, energy prices, and solar energy traces described below.

CDN Trace. We perform experiments using a month-long content delivery network (CDN) dataset from the Akamai CDN provider. The trace contains information about 113k servers geographically distributed across 2691 locations worldwide. It provides the number of servers, capacity of servers, and load information for all the sites at a five-minute granularity. The meta information for each site includes the site's latitude, longitude, city, state, and country.

Carbon Intensity Trace. We use carbon intensity data from ElectricityMaps [10]. The traces provide hourly average carbon intensity information, measured in grams of carbon dioxide equivalent per kilowatt-hour (g·CO_2eq/kWh), for 123 zones worldwide for 2021.

Electricity Prices Trace. We aggregate the electricity price data from multiple sources to include the US Department of Energy [48] and Ember [11] for 2021. The traces provide monthly average electricity prices, measured in ¢/kWh, across different states and countries.

Latency Traces. We utilize latency traces from WonderNetwork [51] which provides round-trip latency between 250 location between July 19th and 20th, 2020.

Solar Trace. We gather solar energy dataset using the PVWatts tool from NREL [8]. It consists of hourly solar energy generation data for a typical meteorological year based on the selected location. We size an individual solar panel for 1kW DC power ratings and scale it to estimate the output of bigger solar panels. Table 4 lists the values of parameters used for the PVWatts trace. We use the default values for all the other parameters.

5.2 Experimental Setup

In our experiments, since different data sources' granularities and lengths do not match, we implement the following processing steps:

(1) We repeat the monthly CDN trace for each month to construct a year-long trace, which enabled us to capture the effects of seasonal variations in carbon intensity and costs.
(2) We consolidate all data centers inside a region (e.g., a state or a country) into a larger data center with the sums of the loads and capacities of the composing data centers and assign them the same carbon intensity and cost. This allowed us to solve the linear program and compute the savings much faster.
(3) We assume that the latency between the clients and their original destinations in the Akamai trace has negligible latency, which enables us to focus on the latency overheads from load shifting.
(4) We assume that data does not change for traces with limited granularity. For example, in the US and Europe, hourly carbon intensity traces are available. At the same time, only monthly energy prices are available. Lastly, we only had a single value for the entire year for some locations in Africa and Asia.
(5) For regions with missing latency data, we use average latency from regions with similar distances from [51].

We implement CDN-Shifter load shifting and capacity shifting policies using Google OR-Tools [36] across different settings. In our load-shifting experiments, we feed our parameters to the solver at a five-minute step. In contrast, we use average quantities within all traces in capacity-shifting experiments. We note that in all experiments, we use latency overhead factor λ and capacity shifting factor γ of 0.1 and 0.01, respectively. Finally, we evaluate the effect of added renewable energy based on total yearly consumption, where we scale the DC system size to match the entire annual consumption in the data center where it is installed.

5.3 Evaluation Metrics

We use three metrics to quantify the benefits and overheads of load and capacity shifting.

Carbon Savings (%). The percentage reduction in operational carbon emissions after spatial load shifting compared to the baseline of no workload migration. We note that negative carbon savings denote cases where emissions increased.

Cost Savings (%). The percentage reduction in operational monetary costs after spatial load shifting compared to the baseline of no workload migration. We note that negative cost savings denote cases where monetary costs increased.

Latency Increase (ms). The increase in latency after spatial load shifting compared to the baseline of no workload migration.

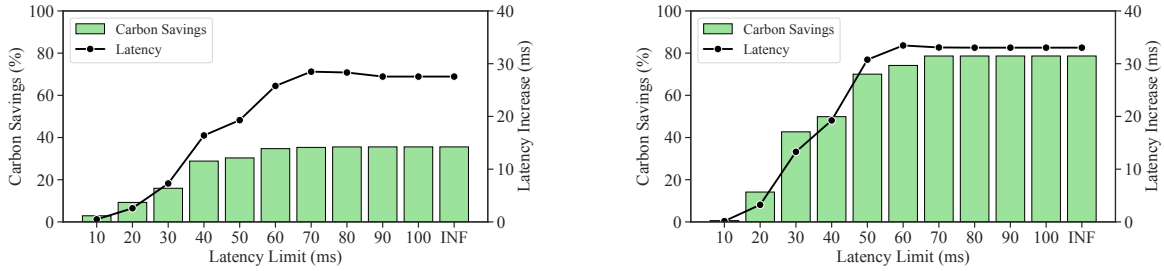
5.4 Effect of Spatial Load Shifting

We start by evaluating the potential of spatial load-shifting within the status quo of the content delivery network, i.e., without capacity shifting or adding renewable. The potential of spatial load shifting depends on three factors: The availability of spare capacity in the network to migrate your workload around, the variations in the carbon intensity and energy prices of various geographically distributed CDN sites, and the latency overheads that users can tolerate. We start by evaluating carbon-aware or cost-aware load shifting individually, and then we look at the trade-offs and methods to co-optimize all operational costs.

5.4.1 Carbon-aware Load Shifting.

The diversity in carbon intensity across edge data centers demonstrates the potential of carbon savings by moving workload from high-carbon locations to low-carbon ones. Figure 7 depicts the carbon savings and latency increases when applying carbon-aware load shifting (by using $\alpha = 1$ in Equation 5) within the USA and Europe. We evaluate the carbon savings under different latency constraints between 10ms and 100ms. We also add a scenario where the load can be migrated anywhere by removing the latency constraints (i.e., $N_i^\delta = N$). As shown, small increases in latency limit yields significant carbon savings. For instance, a 30ms limit yields 15.9%, 42.6%, and 29.1% savings across the US, Europe, and worldwide, respectively.

Figure 7 also highlights the potential savings across the US (see Figure 7a) and Europe (see Figure 7b), where carbon-aware load shifting is able to produce more carbon savings



(a) US **(b) Europe**

Figure 7: Carbon savings (%) and latency increases (ms) in the (a) USA and (b) Europe using carbon-aware spatial load shifting.

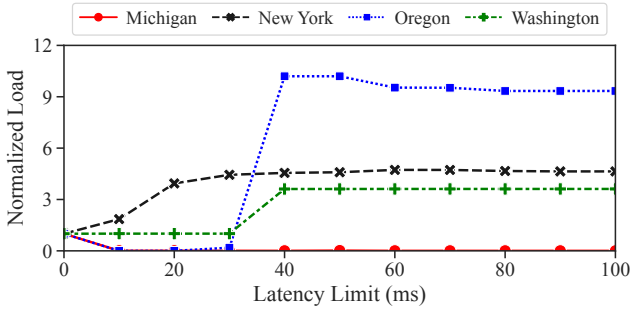


Figure 8: Normalized load w.r.t. carbon-agnostic across latency limits using carbon-aware load shifting in the US.

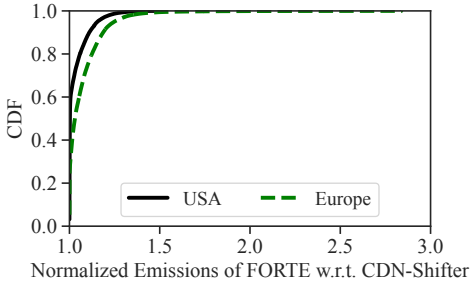


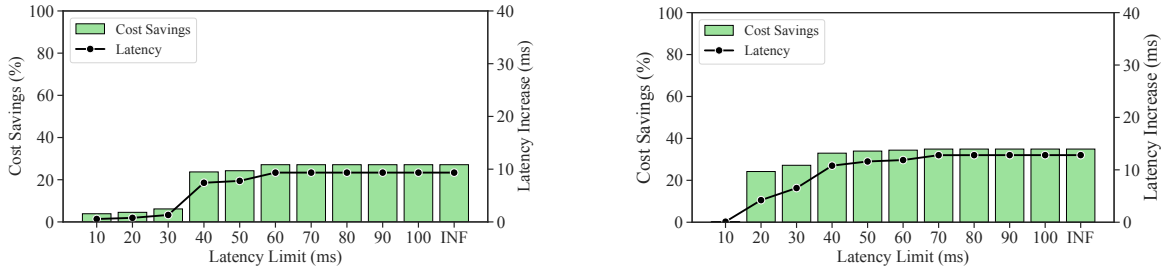
Figure 9: Comparing carbon emissions of CDN-Shifter with FORTE [12].

in Europe, where maximum carbon savings are 35.5% and 78.6%, and 61.7% for US, Europe, and worldwide, respectively. Moreover, the figure shows that carbon savings typically plateau around a latency limit of 60ms, where further increases in the latency limit do not yield further savings. This occurs as all the greenest edge data centers with available capacity have been filled.

Figure 8 illustrate how carbon-aware load shifting alters demand in the US across different latency limits. The figure shows four representative regions that represent different load changes. With an increased latency limit, locations with green energy sources receive more load, while locations with brown regions receive less load. For example, Michigan, a region with an average carbon intensity of 562 g·CO_2eq/kWh, offloads its entire load to other locations starting at a latency limit of 10ms. In contrast, New York receives up to 4.6× its original load. Moreover, some states exhibit nonuniform changes in assigned load. For instance, the total load in Oregon's CDNs decreases and then increases with increases in the latency limit. This is because, with a small latency overhead, Oregon offloads its load to Washington as it has greener energy. Then, once the latency limit increases enough, data centers in Oregon start receiving loads from other brown locations. Finally, the graph highlights an example where the capacity limits the load changes. For example, Washington, the state with the greenest energy, only receives 3.6× its maximum load. Later, we show how capacity shifting can alleviate such limitations, further increasing carbon savings.

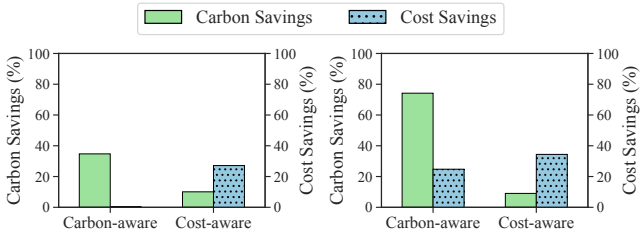
Finally, we evaluate CDN-Shifter's ability to leverage real-time carbon intensity variations across edge data centers to minimize carbon emissions. Figure 9 compares the year long performance of CDN-Shifter to FORTE [12], a state-of-the-art policy that uses static carbon intensity values per region. The figure shows that due to the use of static carbon intensity values FORTE [12] emissions can emit 6% and 3% on average and up to 2.8× and 1.5× more carbon than CDN-Shifter in the worst case for US and Europe, respectively.

Key Takeaway: *Carbon-aware spatial load shifting can result in significant carbon savings of up to 15.9%, 42.6%, and 29.1% for a latency limit of 30ms across the US, Europe, and worldwide, respectively. A latency limit of 60ms yields 97.6% and 94.3% of the maximum carbon savings across the US and Europe, respectively.*



(a) USA **(b) Europe**

Figure 10: Cost savings (%) and latency increases (ms) in the (a) US and (b) Europe using cost-aware spatial load shifting.



(a) US **(b) Europe**

Figure 11: The conflict between carbon-aware and cost-aware shifting in the US and Europe with a latency limit of 60ms.

5.4.2 Cost-aware Load Shifting.

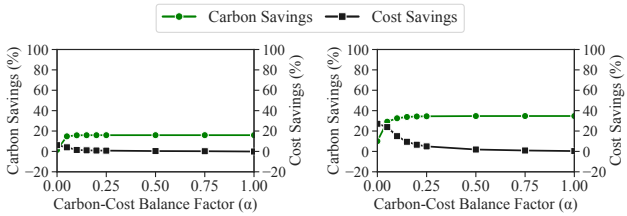
In addition to carbon savings, spatial shifting can be exploited to reduce operational costs. Figure 10 show the cost savings and latency increases when using cost-aware load shifting (i.e., setting $\alpha = 0$ in Equation 5). Similar to the earlier section, we evaluate cost savings under different latency limits and add the scenario with unlimited latency. As expected, cost-aware load shifting derives substantial benefits. For instance, a 30ms limit yields 6.1%, 27%, and 17.1% cost savings across the US, Europe, and worldwide, respectively. Moreover, cost savings do not increase after a latency limit of 60ms, achieving 27.1% and 34.4% within the US and Europe, respectively.

Key Takeaway: *Cost-aware spatial load shifting can result in significant cost savings. A latency limit of 60ms yields 27.1%, 93.9%, and 17.1% cost savings across the US, Europe, and worldwide, respectively.*

5.4.3 Carbon- and Cost-aware Load Shifting.

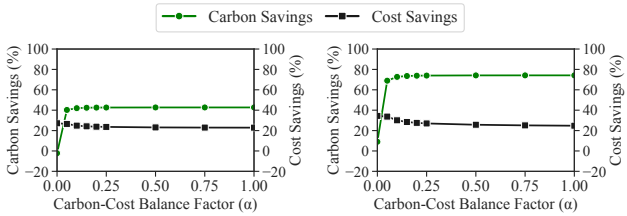
In previous sections, we show how carbon- and cost-aware load shifting can yield significant benefits. One issue with focusing on a single metric (either carbon or cost) is that locations with low carbon intensity may have high energy prices and vice-versa (see Figure 3). Figure 11 demonstrate the breadth of the conflict when we focus on a single metric in the US and Europe, with a 60ms latency. As shown, focusing on a single metric hurts or squanders saving opportunities in the other metric. For example, in Europe, carbon-aware shifting increases operational costs by 16.4%. In addition, although cost-aware scheduling does not increase carbon emissions, it highly limits the possible savings. For example, for a latency limit of 60ms, cost-aware scheduling reduces carbon savings by 71% and 80% across the US and Europe.

To navigate the trade-off between operational emissions and costs, we explore the effect of α, the carbon-cost balance



(a) 30ms **(b) 60ms**

Figure 12: Effect of α on carbon and cost savings in the US, across different latency limits.



(a) 30ms **(b) 60ms**

Figure 13: Effect of α on carbon and cost savings in Europe, across different latency limits.

factor, in the load-shifting decisions in Equation 5. Figures 12 and 13 demonstrate the effect of α in balancing the carbon and cost savings across the US and Europe. As shown, α values between 0.05 and 0.1 depict the balance point where carbon and cost savings are similar. The reason for these small values of α is the difference in magnitude in energy's carbon intensity and costs (See Figure 1 and Figure 2). Figures 12a and 13a depict carbon and cost savings for a latency limit of 30ms, where $\alpha = 0.05$ balances both savings and achieves carbon and cost savings of 29.1% and 23.8% for the US and 40.24% and 26.5% for Europe, respectively.- Finally, it is worth noting that across latency limits, both the carbon and cost savings and the trade-offs between carbon and cost matter, affecting the balance points. For instance, in Europe, the balance points can lead to carbon and cost savings of ~20% and ~35% for latency limits of 30ms and 60ms, respectively.

Key Takeaway: *By carefully choosing α, joint optimization of carbon and cost for load shifting can yield good reductions for both metrics. Specifically, for α ranging from 0.05 to 0.1, CDN-Shifter can achieve carbon and cost savings of 29.1% and 23.8% for the US and 40.24% and 26.5% for Europe, respectively.*

5.5 Effect of Capacity Shifting

When comparing the possible carbon savings (Figure 4) and actual carbon savings (Figure 7a) in the US, it's evident that actual savings for the same latency limits are much smaller. For example, for a latency limit of 100ms, actual savings are smaller by 56.6%. The reason for this is that although some regions hold the potential to reduce carbon emissions, actual load shifting is often bounded by the capacity limits in locations where energy is cheap and green. In this section, we evaluate the potential of capacity shifting in increasing carbon and cost reductions.

Figure 14 shows the effect of carbon-aware capacity shifting across the US (i.e., setting $\alpha = 1$ in Equation 13). The figure compares the carbon savings of load shifting with capacity and load shifting using an expansion factor ψ of 1.5, i.e., capacity can increase by a maximum of 50%. As shown, as the latency limit increases, the role of capacity shifting becomes more apparent. For example, capacity shifting can increment carbon savings by 18.4% and 26.4% for latency limits of 30ms and 60ms, respectively. We note that similar to load shifting, capacity shifting also exhibits diminishing returns.

Table 5 shows how capacity is shifted between regions and the final capacity in the US for the capacity shifting experiment in Figure 14. The table lists examples of areas that gained capacity, such as Washington and New York, where their final capacity increased by 50%. The table also shows examples of places that distributed some of its capacity.

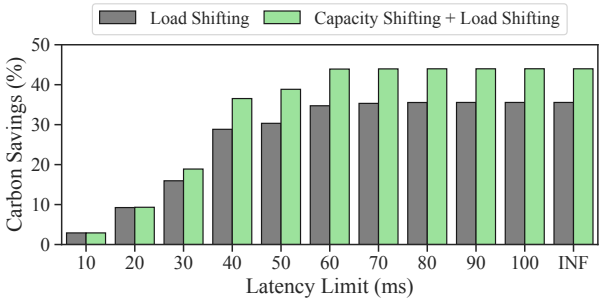


Figure 14: Carbon-aware capacity shifting in the USA using ψ=1.5.

Table 5: Capacity shifting examples between sources (rows) and destinations (columns) from Figure 14.*

State	California	Colorado	New York	Washington
California	76.6%	-	12.4%	9.2%
Colorado	-	48.3%	51.6%	-
New York	-	-	100%	-
Washington	-	-	-	100%
Final Capacity (%)	76.6%	48.3%	150%	150%

* Numbers do not match up as we omitted some sources and destinations.

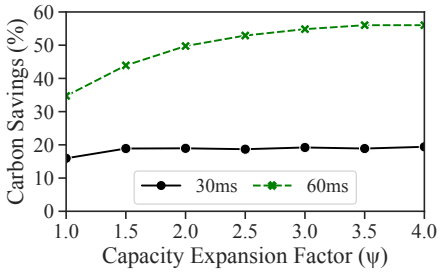
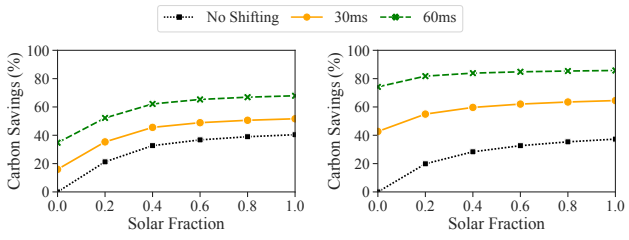


Figure 15: The effect of capacity expansion factor for carbon savings in the US across latency limits.

For instance, California has given out 23.4% of its capacity to regions such as New York and Washington. It is worth noting that, although California has lower carbon intensity than New York, it gave away some of its capacity as it was over-provisioned, with a peak utilization of 30.5%.

To understand the full potential of capacity shifting, Figure 15 evaluates the effect of the capacity expansion factor in the US, using latency limits of 30ms and 60ms. As shown, increments in capacity expansions can increase carbon savings by up to 61.2%. Capacity expansions yield diminishing returns as locations start being short on the capacity they will have to distribute. Moreover, results show that the latency limit may reduce the benefits of spatial shifting, whereas



(a) US **(b) Europe**

Figure 16: Benefits of adding solar energy with and without spatial shifting in the US and Europe.

the results show that expanding the capacity is only effective when the latency limit is high enough. Finally, we note that cost-aware capacity shifting and capacity shifting in Europe exhibit similar performance but are omitted for space limitations.

Key Takeaway: *Capacity shifting overcomes the limitations of load shifting by relocating excess capacity to regions with low carbon and energy costs. In doing so, capacity shifting can increase carbon savings by up to 61.2%.*

5.6 Effect of adding Solar Energy

In the previous section, we considered scenarios where CDNs are only subject to the grid's carbon intensity. However, cloud and CDN providers often use local renewables to reduce operational emissions and costs. In this section, we evaluate the efficacy of adding renewables, specifically solar energy, and how load shifting can amplify the benefits of the added renewables. To model the use of locally available solar energy at each data center during the load shifting optimization, we modified the CDN dataset as follows: at each timeslot t, for each data center, we calculated how much of its compute capacity could be powered by the available local solar energy. We call this capacity C_{solar}, while the remaining capacity of the data center is C_{grid}. We then split the data center into two "virtual" data centers with the exact geographic location, with C_{solar} and C_{grid} as their respective capacities. The virtual data center powered by solar power has a carbon intensity and energy price of 0. In contrast, the virtual data center powered by the grid has the same carbon intensity and energy cost as the original data center. We also split the load of the original data center into two "virtual" data centers. Finally, we run the load-shifting optimization model explained previously. This setup allows us to shift load to data centers with more solar power than they need to power their local load.

Figure 16 shows the carbon saving from solar energy with and without carbon-aware spatial load shifting. We increase the amount of solar power as a fraction of the annual energy consumption, where we scale the deployed DC system to cover the expected yearly demand. Overall, adding local solar capacity without load shifting can reduce carbon emissions by up to 40.4% and 37.3% across the US and Europe. However, the benefits of added solar energy quickly reduce after 40%. For instance, compared to matching 40% of the power consumption with renewables, increasing the solar matching to 100% only adds an extra 7.7% and 8.9% carbon savings for the US and Europe, respectively. One reason for this is that the added solar energy doesn't reduce the emissions of workloads running at night. Another reason is that with increased provisioned solar, the possibility of having excess energy than demand also increases.

One solution to increase the benefits of added renewables is augmenting them with spatial shifting, which can significantly increase carbon savings. For example, in the US, the maximum carbon saving with added renewables can be achieved or surpassed when combining a latency limit of 30ms and only matching 20% of the total energy consumption by solar. The figure shows that when mixing renewable energy with solar energy, carbon savings can reach up to 67.9% and 85.7% across the US and Europe, respectively. Finally, we note that adding renewables has a similar effect on cost savings, yet it doesn't eliminate the carbon-cost trade-off.

Key Takeaway: *Adding renewable energy falls short in reducing the carbon emissions of CDNs. However, load shifting can proliferate the benefits of renewables by 68% and 130% across the US and Europe.*

5.7 Discussion

We have shown the benefits of CDN-Shifter in minimizing carbon emissions of global CDN networks. Next, we highlight other benefits of the proposed methods and their limitations.

Generalizbility of CDN-Shifter. Load and capacity shifting are widely applicable to other interactive services. One particular scenario that has gained a lot of attention recently is serving AI models. Similar to CDN applications, AI requests are primarily stateless and can be freely shifted. In addition, their high computational costs and loose latency requirements make them a perfect candidate for spatial load shifting. Next, when discussing capacity shifting, we used an abstract cloud model and focused on the benefits of transferring virtual resource capacity between locations to increase the available capacity at locations with greener energy resources. However, capacity shifting is generally applicable to physical resources, and hyper-scale data centers could utilize it to promote the usage of resources at locations where energy is greener or cheaper. One way to implement this is to extend the lifetime of old and error-prone servers by

moving them to locations with greener energy and offering them at discounted prices [27].

Limitations of CDN-Shifter. The use of spatial load shifting can significantly reduce carbon emissions. However, our approach has limitations. Firstly, we assumed that the required content is available at the new location. This assumption is based on the extensive scale of CDN networks and their ability to distribute popular content quickly, but we did not explicitly consider the associated overheads. Secondly, we assumed that all requests are uniform and subject to the same latency constraints. However, it is feasible to enhance our model to accommodate different request types and latency requirements, and we leave such evaluation to future work. Thirdly, we only considered emissions and costs from energy consumption by edge data centers. However, data transmission has nonnegligible energy consumption and costs. Lastly, we did not consider the cost of incorporating renewable energy sources and methods to optimize this process, a complementary issue that has been studied elsewhere [17, 18].

6 RELATED WORK

Earlier work on load shifting have focused on either reducing operational costs [14, 18, 28, 29, 33, 35, 37, 38], reducing energy consumption [31, 32, 40], or carbon emissions [14, 18, 28, 29]. For example, Mathew et al. [33] and Goiri et al. [14] utilize the flexibility of delay-tolerant batch workloads to execute them when renewable energy is available or when grid-supplied energy is cheaper. In contrast to temporal shifting, which does not suit the latency requirements of interactive workloads, spatial load shifting has seen more popularity in reducing operational costs and emissions. For instance, Qureshi et al. [37] showed how the differences in energy prices across states could be used to decrease total operational costs. Moreover, several authors have explored methods to reduce energy consumption and operational costs. For instance, Mathew et al. [31] explored methods to power down idle servers to save energy while adhering to Service Level Availability (SLA) constraints. Lastly, several authors have explored renewable energy's role in reducing carbon emissions in addition to cost and energy savings. For instance, Liu et al. [29] have explored how renewable-aware load shifting can reduce carbon emissions and operational costs.

Previous work on renewable-aware load shifting often assumes that grid-supplied energy is always brown or fixed. However, more recent research has highlighted the diversity in energy's carbon intensity across locations and times [9, 19–21, 30, 45, 46]. For example, Dodge et al. [9] have utilized spatial load shifting to decrease carbon emissions of AI workloads while respecting SLA constraints. In contrast to earlier work, we consider the three-way trade-off between load shifting and implemented large-scale evaluations of spatial load shifting to highlight the potential of such techniques. Perhaps the most relevant work was done by Gao et al. [12], where they highlighted the trade-off. However, they focused on data placement and load shifting, and their evaluation assumed fixed carbon intensity and costs. In contrast, we evaluate the impact of real-time carbon intensity variations. Additionally, we consider new methods to further reduce emissions and costs, such as capacity shifting and adding local renewables.

In addition to load shifting, previous research has considered the effect of resource planning decisions, where cloud or CDN providers can select regions for their new resources based on energy's cost or carbon intensity [13, 24, 26, 39]. In contrast to previous research, we focus on scenarios that do not require adding new resources by redistributing them. Finally, in addition to load shifting, that magnifies the benefits of the added renewables. Researchers have explored the utilization of batteries to save excess energy and use it when renewable energy is not available or insufficient [14, 35]. In this work, we only consider load shifting. Mixing batteries and load-shifting and evaluating how it affects the three-way carbon, cost, and latency trade-off is left for future research.

7 CONCLUSIONS

In this paper, we studied the potential for using spatial workload shifting using geographic load and capacity shifting to reduce the carbon emissions and energy costs of large-scale CDNs. We formulate these problems as optimization problems, considering the three-way trade-off between carbon emissions, cost, and latency while adhering to capacity constraints. We evaluate the proposed method using real-world CDN workloads, carbon intensity, and energy prices from various electricity grids. Our results show the potential of load shifting in decarbonizing CDNs. Specifically, we show that increasing the latency by 60ms can reduce carbon emissions by up to 35.5%, 78.6%, and 61.7% across the US, Europe, and worldwide, respectively. In addition, we show that capacity shifting can increase carbon savings by up to 61.2%. Finally, we analyze the benefits of spatial shifting and show that it increases carbon savings from added solar energy by 68% and 130% in the US and Europe, respectively.

ACKNOWLEDGEMENTS

We thank the SoCC reviewers and our shepherd, Jelle Hellings, for their valuable comments, which improved the quality of this paper, and electricityMap for access to their carbon-intensity data. This research is supported by NSF grants 23091241, 2213636, 2211302, 2211888, 1763617, 2105494, 2325956, DoE grant DE-EE0010143, and VMware.

REFERENCES

[1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 118–132.

[2] João Aires, Paulo Duarte, Bruno Parreira, and Sérgio Figueiredo. 2020. An orchestrated 5G-enabled deep vCDN system. *Internet Technology Letters* 3, 6 (2020), e189.

[3] Akamai. 2023. Akamai CDN Network Deployment: Facts and Figures. https://www.akamai.com/company/facts-figures.

[4] Srini Bangalore, Arjita Bhan, Andrea Del Miglio, Pankaj Sachdeva, Vijay Sarma, Raman Sharma, and Bhargs Srivathsan. 2023. Investing in the Rising Data Center Economy. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy.

[5] Noman Bashir, David Irwin, Prashant Shenoy, and Abel Souza. 2022. Sustainable Computing – Without the Hot Air. In *Proceedings of the First Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon)*.

[6] California Independent System Operator (CAISO). 2016. What the duck curve tells us about managing a green grid.

[7] Wesley J Cole, Danny Greer, Paul Denholm, A Will Frazier, Scott Machen, Trieu Mai, Nina Vincent, and Samuel F Baldwin. 2021. Quantifying the challenge of reaching a 100% renewable energy power system for the United States. *Joule* 5, 7 (2021), 1732–1748.

[8] Aron P Dobos. 2014. *PVWatts Version 5 Manual*. Technical Report. National Renewable Energy Laboratory (NREL).

[9] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1877–1894.

[10] Electricity Maps. 2024. https://www.electricitymap.org/.

[11] Ember. 2024. Monthly Electricity Data. https://ember-climate.org/.

[12] 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)*. 211–222. https://doi.org/10.1145/2342356.2342398

[13] Daniel Gmach, Jerry Rolia, Cullen Bash, Yuan Chen, Tom Christian, Amip Shah, Ratnesh Sharma, and Zhikui Wang. 2010. Capacity Planning and Power Management to Exploit Sustainable Energy. In *2010 International Conference on Network and Service Management*. 96–103. https://doi.org/10.1109/CNSM.2010.5691329

[14] I Goiri, W Katsak, K Le, T Nguyen, and R Bianchini. 2013. Parasol and GreenSwitch: Managing Datacenters Powered by Renewable Energy. In *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*.

[15] Google. 2024. *Cloud CDN overview*. https://cloud.google.com/cdn/docs/overview

[16] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool. In *ISCA*.

[17] Vani Gupta, Prashant Shenoy, and Ramesh K. Sitaraman. 2018. Efficient Solar Provisioning for Net-Zero Internet-Scale Distributed Networks. In *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*. 372–379. https://doi.org/10.1109/COMSNETS.2018.8328221

[18] Vani Gupta, Prashant Shenoy, and Ramesh K Sitaraman. 2019. Combining Renewable Solar and Open Air Cooling for Greening Internet-Scale Distributed Networks. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. 303–314.

[19] Walid A. Hanafy, Roozbeh Bostandoost, Noman Bashir, David Irwin, Mohammad Hajiesmaili, and Prashant Shenoy. 2023. The War of the Efficiencies: Understanding the Tension between Carbon and Energy Optimization. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems* (Boston, MA, USA) *(HotCarbon '23)*. Article 19, 7 pages. https://doi.org/10.1145/3604930.3605709

[20] Walid A. Hanafy, Qianlin Liang, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing Carbon-Efficiency. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 3, Article 57 (December 2023), 28 pages. https://doi.org/10.1145/3626788

[21] Walid A. Hanafy, Qianlin Liang, Noman Bashir, Abel Souza, David Irwin, and Prashant Shenoy. 2024. Going Green for Less Green: Optimizing the Cost of Reducing Cloud Carbon Emissions. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS '24)*. 479–496. https://doi.org/10.1145/3620666.3651374

[22] Ronald Huisman, Christian Huurman, and Ronald Mahieu. 2007. Hourly Electricity Prices in Day-Ahead Markets. *Energy Economics* 29, 2 (2007), 240–248. https://doi.org/10.1016/j.eneco.2006.08.005

[23] Google Inc. 2023. Building a Carbon-free Future for All. https://sustainability.google/commitments/carbon/.

[24] Hatem Khedher, Emad Abd-Elrahman, Hossam Afifi, and Michel Marot. 2017. Optimal and Cost Efficient Algorithm for Virtual CDN Orchestration. In *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*. 61–69. https://doi.org/10.1109/LCN.2017.115

[25] Navaneeth Krishna. 2021. CDN. https://almanac.httparchive.org/en/2021/cdn.

[26] Dan Liao, Gang Sun, Guanghua Yang, and Victor Chang. 2018. Energy-efficient virtual content distribution network provisioning in cloud-based data centers. *Future Generation Computer Systems* 83 (2018), 347–357. https://doi.org/10.1016/j.future.2018.01.057

[27] Yejia Liu, Pengfei Li, Daniel Wong, and Shaolei Ren. 2024. Geographical Server Relocation: Opportunities and Challenges. In *Proceedings of HotCarbon'24*. https://hotcarbon.org/assets/2024/pdf/hotcarbon24-final20.pdf Accessed: 2024-10-08.

[28] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, M. Marwah Z. Wang, and C. Hyser. 2012. Renewable and Cooling Aware Workload Management for Sustainable Data Centers. In *Proceedings of ACM Sigmetrics*.

[29] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew. 2012. Greening Geographical Load Balancing. In *Preprint. Extension of a paper that appeared in ACM Sigmetrics, 2011*.

[30] 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, HotCarbon 2023, Boston, MA, USA, 9 July 2023*. 6:1–6:6.

[31] Vimal Mathew, Ramesh K. Sitaraman, and Prashant Shenoy. 2012. Energy-aware Load Balancing in Content Delivery Networks. In *Proceedings IEEE INFOCOM*. 954–962. https://doi.org/10.1109/INFCOM.2012.6195846

[32] Vimal Mathew, Ramesh K. Sitaraman, and Prashant Shenoy. 2013. Energy-Efficient Content Delivery Networks using Cluster Shutdown. In *2013 International Green Computing Conference Proceedings*. 1–10. https://doi.org/10.1109/IGCC.2013.6604510

[33] Vimal Mathew, Ramesh K Sitaraman, and Prashant Shenoy. 2014. Reducing Energy Costs in Internet-Scale Distributed Systems Using Load

Shifting. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*. IEEE, 1–8.

[34] E. Nygren, R.K. Sitaraman, and J. Sun. 2010. The Akamai Network: A Platform for High-performance Internet Applications. *ACM SIGOPS Operating Systems Review* (2010).

[35] Darshan S Palasamudram, Ramesh K Sitaraman, Bhuvan Urgaonkar, and Rahul Urgaonkar. 2012. Using Batteries to Reduce the Power Costs of Internet-scale Distributed Networks. In *Proceedings of the Third ACM Symposium on Cloud Computing*. 1–14.

[36] Laurent Perron and Vincent Furnon. 2024. *Google OR-Tools v9.10*. Google. https://developers.google.com/optimization/

[37] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. 2009. Cutting The Electric Bill for Internet-Scale Systems. In *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*. ACM, 123–134.

[38] H. Rahul, M. Kasbekar, R. Sitaraman, and A. Berger. 2006. Towards Realizing the Performance and Availability Benefits of a Global Overlay Network. In *Proc. of Passive and Active Measurement Conference*. Citeseer.

[39] Chuangang Ren, Di Wang, Bhuvan Urgaonkar, and Anand Sivasubramaniam. 2012. Carbon-Aware Energy Capacity Planning for Datacenters. In *2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. 391–400. https://doi.org/10.1109/MASCOTS.2012.51

[40] Nishanth Sastry and Jon Crowcroft. 2010. SpinThrift: saving energy in viral workloads. In *Proceedings of the First ACM SIGCOMM Workshop on Green Networking* (New Delhi, India) *(Green Networking '10)*. 69–76. https://doi.org/10.1145/1851290.1851305

[41] Fred C Schweppe, Michael C Caramanis, Richard D Tabors, and Roger E Bohn. 2013. *Spot Pricing of Electricity*. Springer Science & Business Media.

[42] David Shepardson and Nandita Bose. 2019. Reuters, Amazon Vows to be Carbon Neutral by 2040, buying 100,000 Electric Vans. https://www.reuters.com/article/us-amazon-environment/amazon-vows-to-be-carbon-neutral-by-2040-buying-100000-electric\-vans-idUSKBN1W41ZV.

[43] Brad Smith. 2020. Microsoft will be carbon negative by 2030. https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/

[44] Abel Souza, Noman Bashir, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. 2023. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications. In *ASPLOS*.

[45] Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2023. CASPER: Carbon-Aware Scheduling and Provisioning for Distributed Web Services. In *Proceedings of the 14th International Green and Sustainable Computing Conference (IGSC), Toronto, ON, Canada*.

[46] 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 (EuroSys '24)*. Association for Computing Machinery, New York, NY, USA, 924–941. https://doi.org/10.1145/3627703.3650079

[47] Akamai Technologies. 2024. Akamai Traffic Map. https://www.akamai.com/internet-station/traffic-map. Accessed January 2024.

[48] U.S. Energy Information Administration (EIA). 2024. http://www.eia.doe.gov.

[49] WattTime. 2022. WattTime. https://www.watttime.org/.

[50] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In *Proceedings of the 22nd International Middleware Conference (Middleware)*. 260–272.

[51] WounderNetwork. [n. d.]. Global Ping Statistics. https://wondernetwork.com/pings. Accessed: 2024-5-30.