Title

Towards Environmentally Equitable AI via Geographical Load Balancing

Authors

Li, Pengfei

Yang, Jianyi

Wierman, Adam

et al.

Publication Date

2023-06-20

Towards Environmentally Equitable AI via Geographical Load Balancing

Pengfei Li
UC Riverside
pli081@ucr.edu

Jianyi Yang
UC Riverside
jyang239@ucr.edu

Adam Wierman
Caltech
adamw@caltech.edu

Shaolei Ren[1]
UC Riverside
sren@ece.ucr.edu

Abstract

Fueled by the soaring popularity of large language and foundation models, the accelerated growth of artificial intelligence (AI) models' enormous environmental footprint has come under increased scrutiny. While many approaches have been proposed to make AI more energy-efficient and environmentally friendly, environmental inequity — the fact that AI's environmental footprint can be disproportionately higher in certain regions than in others — has emerged, raising social-ecological justice concerns. This paper takes a first step toward addressing AI's environmental inequity by balancing its regional negative environmental impact. Concretely, we focus on the carbon and water footprints of AI model inference and propose equity-aware geographical load balancing (GLB) to explicitly address AI's environmental impacts on the most disadvantaged regions. We run trace-based simulations by considering a set of 10 geographically-distributed data centers that serve inference requests for a large language AI model. The results demonstrate that existing GLB approaches may amplify environmental inequity while our proposed equity-aware GLB can significantly reduce the regional disparity in terms of carbon and water footprints.

Source code: `https://github.com/Ren-Research/Environmentally-Equitable-AI`

1 Introduction

Building on the advances in deep neural networks, artificial intelligence (AI) has become an indispensable powerhorse for enabling scientific breakthroughs, accelerating business growth, and addressing global challenges in numerous domains of critical needs [1, 2]. The success of AI relies heavily on computationally-intensive calculations to learn useful information from data during training and provide insightful predictions during inference. As such, AI models, especially large generative models like GPT-3 [3], are typically trained on large clusters of power-hungry servers that may each have multiple graphic processing units (GPUs) and are housed in warehouse-scale data centers. Moreover, for inference, AI models are often deployed in geographically distributed data centers to server users with low transmission latency.

Consequently, the exponentially growing demand for AI has created an enormous appetite for energy as well as a negative impact on the environment [3–8]. For example, putting aside the environmental toll of chip manufacturing (e.g., raw material extraction and toxic chemicals) [9–11] and the noise pollution of running AI servers [12], training a large language model like GPT-3 and LaMDA can easily consume hundreds of megawatt-hour of electricity, generate many tonnes of carbon emissions, and evaporate hundreds of thousands of liters of clean freshwater for cooling [8, 13, 14]. In fact, even after adopting the industry's best practices to curb AI's resource usage, the overall energy consumption by AI models at Google has taken up ∼10-15% of its data centers' energy consumption [4]. Crucially, in addition to their impacts on the global climate, AI's environmental footprint also has significant local and regional impacts. Elevated carbon emissions have localized social costs [15] and may increase local ozone, particulate matter, and premature mortality [16]; electricity generation, especially when burning fuels, produces local air pollutants, discharges pollution such as thermal pollution into water bodies, and generates solid wastes (possibly including hazardous wastes) [17]; and staggering water consumption, both directly for on-site cooling and indirectly for off-site electricity generation, can further stress the already-limited local freshwater resources and even worsen extended megadroughts in regions like Arizona [14, 18].

Fueled by the soaring popularity of large language and foundation models, the accelerated growth of AI's environmental footprint has come under increased scrutiny recently [19, 20]. To make AI more energy-efficient and environmentally friendly, research studies have pursued a variety of approaches, including computationally efficient training and inference [21, 22], energy-efficient GPU and accelerator designs [4, 23, 24],

[1]Corresponding author: Shaolei Ren

carbon-aware task scheduling [7, 20], green cloud infrastructures [25–27], sustainable AI policy recommendations [10,19], among others. As supply-side solutions, data center operators have also increasingly adopted carbon-free energy such as solar and wind power, (partially) powering AI servers and lowering carbon emissions [20,28,29]. Additionally, to reduce on-site water consumption and mitigate the stress on already-limited freshwater resources, climate-conscious cooling system designs (e.g., using air-side economizers if the climate condition permits) have recently seen an uptick in the data center industry [30, 31].

While existing efforts are encouraging, a worrisome outcome — environmental inequity — has unfortunately emerged. That is, AI's environmental footprint is disproportionately higher in certain regions than in others, potentially exacerbating other unintended social-ecological consequences [32]. For example, a data center's on-site cooling water usage effectiveness (WUE, the ratio of water consumption to IT energy consumption) highly depends on the outside temperature [33] — while it can stay below 1.0 L/kWh for data centers located in cooler climates, the monthly average WUE can be as high as 9.0 L/kWh in the summer in drought-stricken Arizona [34]. Likewise, there exists a significant regional difference in terms of the carbon efficiency — as of 2020, only 4% of the energy for Google's data center in Singapore is carbon-free, whereas this number goes up to 94% in Finland [4], creating a 23$\times$ disparity. Thus, as a result of such regional differences, certain data center locations are severely "disadvantaged" and more negatively impacted by the environmental toll of AI. Further compounded by enduring socioeconomic disparities and even potentially amplified by existing data center scheduling algorithms, environmental inequity of AI can pose critical business risks and hence needs to be properly reconciled.

Indeed, while it is crucial to mitigate AI's algorithmic unfairness against disadvantaged individuals or user groups [35–39], addressing its environmental inequity is also increasingly important and becoming integral to responsible AI. For example, in the first-ever global agreement to ensure healthy development of AI, the United Nations Educational, Scientific and Cultural Organization (UNESCO) recommends that "AI should not be used" if it creates "disproportionate negative impacts on the environment" [40]. Among all the environmental-related topics, Meta ranks environmental *justice* as the most critical one with the greatest impact on its business risks and opportunities [29]. More recently, studies have also emerged to suggest new regulations pertinent to AI's growing environmental footprint [19], and holistic assessment of AI as social-ecological-technological systems using available tools from environmental justice [41]. In a different but relevant context, the disproportionate environmental impacts (e.g., air pollution and water consumption) of energy-consuming digital asset operations have already raised environmental justice concerns which, according to the recommendation of the U.S. White House Office of Science and Technology, need to be addressed as a priority to support responsible development [42].

In this paper, we take a first step to address the emerging environmental inequity of AI by balancing its negative environmental impact across geographically-distributed data centers. More concretely, we focus on the carbon and water footprints of AI model inference and dynamically schedule users' inference requests (also referred to as *workloads* in this paper) using equity-aware geographical load balancing (GLB). To mitigate environmental inequity, the key novelty of our GLB approach is that we augment the traditional cost-saving objective by explicitly including minimization of the most significant negative environmental impacts among all the data centers. We also extend our GLB problem formulation to consider more advanced settings where there is on-site carbon-free energy available to power the AI workloads and where we can further exploit AI's energy-accuracy flexibility by dynamically choosing one or more AI models to serve the workloads.

To empirically evaluate our proposed equity-aware GLB, we run trace-based simulations by considering a set of 10 geographically-distributed data centers that serve inference requests for a large language AI model. Our results demonstrate that the proposed equity-aware GLB can significantly reduce the carbon and water footprints in the most disadvantaged region. In stark contrast, existing carbon- and water-saving GLB approaches may even amplify environmental inequity.

In summary, our work is the first study to advance AI's environmental equity via GLB, connecting research across data center scheduling, sustainable AI, and equitable AI. It highlights the need and great potential of equity-aware GLB to address AI's emerging environmental inequity.

2 Related Works

Our work contributes to the literature on GLB for cloud computing and data centers [26, 27, 33, 43–54]. Prior studies focus on reducing the total energy cost, carbon footprint, water footprint, or a weighted combination

of these metrics; ignoring the potential for regional disparities. We show in this paper that existing GLB algorithms can potentially amplify environmental inequity by further exploiting already vulnerable regions. For example, GLB algorithms that aggressively exploit lower electricity prices [50,53] and/or more renewables [51, 52] may schedule more workloads to data centers (located in, for example, Arizona) that are extremely water-stressed; thus adding a disproportionately high pressure to local water systems.

Sustainable AI has received a significant amount of attention in recent years [3–7,13,55]. To make AI more energy-efficient and sustainable, a variety of approaches have been explored and studied, including computationally efficient training and inference [21, 22], energy-efficient GPU and accelerator designs [4, 23, 24], carbon-aware task scheduling [7, 20], green cloud infrastructures [25–27, 56], among others. While they are useful for overall sustainability, these studies do not address the emerging environmental equity among different regions for deploying AI services. Additionally, they have mostly focused on carbon footprint, neglecting other crucial environmental footprints, e.g., water footprint [14, 28, 29, 57]. In contrast, we holistically consider both carbon and water footprints and make novel contributions to sustainable AI from the perspective of environmental equity.

There also exist non-*computational* approaches to improving AI's environmental sustainability. For example, data center operators have increasingly adopted carbon-free energy such as solar and wind power to lower AI's carbon emissions [20, 28, 29, 57]. To cut on-site potable water consumption and mitigate the stress on already-limited freshwater resources, climate-conscious cooling system designs (e.g., air-side economizers and purifying non-potable water) have recently seen an uptick in the data center industry [30,31]. These non-computational approaches alone are typically not the most effective solution to sustainable AI, and must be designed in conjunction with computational approaches (e.g., workload scheduling) [27,44,58,59]. As such, our study of equity-aware GLB can inform the planning of on-site carbon-free energy and cooling system renovation projects to better achieve social and environmental justice.

Equity and fairness are crucial considerations for the success of AI. The existing research in this space has predominantly focused on mitigating prediction unfairness against disadvantaged individuals and/or groups under a variety of settings [35–39, 60–66]. Our work on environmental equity adds a unique dimension of fairness and greatly complements the existing rich body of research, collaboratively and holistically building equitable and socially-responsible AI.

3 Problem Formulation

We consider a pre-trained AI model (e.g., large language model) that is deployed for inference services over a set $\mathcal{N} = \{1, \cdots, N\}$ of geographically distributed data centers to serve users in different regions. There are a set $\mathcal{J} = \{1, \cdots, J\}$ of front-end traffic gateways that aggregate users' requests from their respective surrounding areas and assign the requests to data centers, which is also referred to as geographical load balancing (GLB) in the literature [51]. The GLB decisions are made in a time-slotted manner over a total of T time slots. In practice, each time slot can range from a few minutes to about an hour, depending on how frequently the decisions are updated. We also interchangeably use "workloads" and "requests" when referring to users' demand for the AI model inference service.

Each data center houses a cluster of servers (typically each equipped with multiple GPUs) to host AI models for inference. For the ease of presentation, we assume a homogeneous AI model on all the servers, while the extension to heterogeneous AI models with different model sizes is considered in Section 4.2.2. During each time slot, the maximum service capacity for the AI model inference is M_i for data center i. We use $\lambda_{j,t}$ to denote the total amount of workloads arriving at gateway j at time t, and $x_{ij}(t) \geq 0$ to represent the GLB decision (i.e., the load assigned to data center i from gateway j). For the convenience of presentation, we also use $x(t) = \{x_{i,j}(t) \,|\, i \in \mathcal{N}, j \in \mathcal{J}\}$ as the collection of all the GLB decisions at time t.

The total load assigned to data center i is $\sum_{j\in\mathcal{J}} x_{ij}(t) \leq M_i$ at time t, thus resulting in a total server energy consumption of $e_i(x(t))$ which is an increasing function of $\sum_{j\in\mathcal{J}} x_{ij}(t)$. For example, $e_i(x(t))$ can be expressed as $e_i(x(t)) = \rho_{i,t}\bar{E}_{i,s} + \frac{\sum_{j\in\mathcal{J}} x_{ij}(t)}{M_i} \cdot \bar{E}_{i,d}$ where $\bar{E}_{i,s}$ is the server cluster's static/idle energy even when no workload is processed in data center i, $\bar{E}_{i,d}$ is the cluster's dynamic energy consumed when only processing workloads, $\frac{\sum_{j\in\mathcal{J}} x_{ij}(t)}{M_i}$ is the cluster-level utilization, and $\frac{\sum_{j\in\mathcal{J}} x_{ij}(t)}{M_i} \leq \rho_{i,t} \leq 1$ indicates how well the cluster is right-sized in proportion to the workloads (i.e., $\rho_{i,t} = \frac{\sum_{j\in\mathcal{J}} x_{ij}(t)}{M_i}$ means the cluster is perfectly sized to the workloads by turning off unused servers, while $\rho_{i,t} = 1$ means the servers are always kept on

regardless of the assigned workloads).

Next, we model the energy cost, carbon footprint, and water footprint in terms of the GLB decisions.

Energy cost. Suppose that the electricity price and power usage effectiveness (PUE, which accounts for non-IT energy consumption such as cooling systems and power distribution losses) are $p_{i,t}$ and $\gamma_{i,t}$ for data center i at time t, respectively. Then, the total energy cost at time t can be written as

$$g_t(x(t)) = \sum_{i \in \mathcal{N}} p_{i,t} \gamma_{i,t} e_i(x(t)). \tag{1}$$

Note that, if the AI model inference service is run on virtual machine (VM) instances rented from public cloud providers, the electricity price $p_{i,t}$ becomes the VM price subject to the VM instance type and $g_t(x(t)) = \sum_{i \in \mathcal{N}} p_{i,t} e_i(x(t))$ is the total VM rental cost at time t where $e_i(x(t))$ represents the number of VM instances rented to process the assigned workloads in location i.

Carbon footprint. The carbon footprint of AI model inference is embedded in the generation of electricity using carbon-intensive fuels such as coal [8, 11, 52]. By denoting the carbon efficiency as $\alpha_{i,t}$ with a unit of gram/kWh, we have the following carbon footprint for data center i at time t:

$$c_{i,t}(x(t)) = \alpha_{i,t} \gamma_{i,t} e_i(x(t)) \tag{2}$$

The carbon efficiency $\alpha_{i,t}$ can be obtained by querying the local utility or averaging the carbon efficiency of the grid's fuel mix.

Water footprint. To serve AI model inference, data centers consume clean freshwater both directly and indirectly [14, 30, 67, 68]. The direct water consumption comes from the cooling system to keep servers from overheating. Specifically, data centers commonly use cooling towers as the heat rejection mechanism due to their energy efficiency and applicability to a wide range of weather conditions, but a large amount of water is evaporated into the outside environment (i.e., not returning to the source and instead entering the global water ecosystem) and hence considered "consumed" [30]. For example, depending on the outside wet-bulb temperature, a cooling tower typically consume 1∼4 liters of water (up to 9 liters of water in the summer) for each kWh server energy [34]. Importantly, the vast majority of the cooling water supply is drinking-grade (e.g., nearly 90% for Google's U.S. data centers in 2021 [30]). Although air-side economizers (i.e., directly using outside air to cool down servers) can be used save water if the climate condition is sutiable, water is still needed when the outside temperature is high and/or the humidity is low — Meta's state-of-the-art cooling systems use an average of 0.26 liters of water for each kWh server energy across its global data center fleet in 2021 [29]. Additionally, electricity generation incurs significant water footprint (e.g., coal and nuclear power plants require a large volume of water consumption for cooling), with the U.S. national average electricity water intensity at around 1.8 L/kWh (excluding hydropower) [14,18,69]. Thus, the same as carbon footprint, AI models are also accountable for indirect water footprint. By combining both direct and indirect water consumption, the water footprint for data center i at time t is

$$w_{i,t}\left(x(t)\right) = \left[\epsilon_{i,t} + \beta_{i,t}\gamma_{i,t}\right] \cdot e_i\left(x(t)\right), \tag{3}$$

where $\epsilon_{i,t}$ is the direct water usage effectiveness (WUE) for on-site cooling, $\beta_{i,t}$ is the indirect WUE for off-site electricity generation, and $\gamma_{i,t}$ is the PUE. Note that the direct WUE is defined as the ratio of water consumption to IT server energy consumption [70], and hence we do not need to multiply $\gamma_{i,t}$ when calculating the direct water consumption. In practice, the direct WUE $\epsilon_{i,t}$ heavily depends on the outside temperature, and hence can be modeled as a time-varying function in terms of the outside weather condition [14,34]. Like the carbon efficiency, the indirect WUE $\beta_{i,t}$ measures the water consumption per kWh electricity generation and can be calculated by averaging over the water intensity of different energy fuels [33]. Note that, as a large bulk of the on-site water cost is the fixed connection charge based on the maximum water consumption rate, the cost for the actual on-site water usage is typically much smaller compared to the energy cost. Thus, we exclude the actual monetary water cost from our problem formulation.

4 Geographical Load Balancing for Environmentally Equitable AI

To make AI environmentally equitable, we propose to mitigate the disproportionality of AI's negative environmental impacts on different regions by optimally balancing AI workloads among geographically distributed data centers.

4.1 Basic Setting

We first consider the setting in which all the data centers are powered by grid electricity and the inference workloads are served by homogeneous AI models. While addressing AI's environmental inequity, our goal is not to blindly *equalize* its regional environmental footprint, which, as similarly observed in the context of mitigating AI's algorithmic unfairness [35], may artificially elevate the environmental footprints in those otherwise advantaged regions and provide a false sense of fairness. Instead, we adopt the notion of *minimax* fairness [35, 63, 71] and exploit the power of GLB as a software-based approach to explicitly minimize AI's environmental impact on the most disadvantaged region.

Mathematically, we augment the cost-saving objective by including minimization of the greatest environmental impacts among all the data centers. Our equity-aware GLB problem is formulated as follows:

$$\min_{x(t)} \sum_{t=1}^{T} g_t(x(t)) + \mu_c \cdot \max_{i\in\mathcal{N}} \left[\mathcal{H}_{i,c}\left(\sum_{t=1}^{T} c_{i,t}(x(t))\right)\right] + \mu_w \cdot \max_{i\in\mathcal{N}} \left[\mathcal{H}_{i,w}\left(\sum_{t=1}^{T} w_{i,t}\left(x(t)\right)\right)\right], \tag{4a}$$

$$s.t. \quad x_{i,j}(t) = 0, \ \text{ if } B_{i,j} = 0, \quad \forall\, i \in \mathcal{N}, j \in \mathcal{J}, t = 1, \cdots, T, \tag{4b}$$

$$\sum_{j\in\mathcal{J}} x_{i,j}(t) \le M_i, \quad \forall\, i \in \mathcal{N}, t = 1, \cdots, T, \tag{4c}$$

$$\sum_{i\in\mathcal{N}} x_{i,j}(t) = \lambda_{i,j}, \quad \forall\, j \in \mathcal{J}, t = 1, \cdots, T, \tag{4d}$$

where the assignment condition $B_{i,j} = 0$ indicates that the workloads cannot be assigned from gateway j to data center i (due to, e.g., latency constraints or data sovereignty regulations) and hence enforces $x_{i,j} = 0$ in (4b), the constraint (4c) means that the total workloads assigned to a data center cannot exceed its processing capacity, and the constraint (4d) requires that all workloads arriving at a gateway be assigned to data centers. In the optimization objective (4a), the monotonically-increasing functions $\mathcal{H}_{i,c}()$ and $\mathcal{H}_{i,w}()$ quantify the negative environmental impacts of AI on data center i due to its carbon footprint and water footprint, respectively, and can be specified based on the local environment assessment.

Using a linear function $\mathcal{H}_{i,w}\left(\sum_{t=1}^{T} w_{i,t}\left(x(t)\right)\right) = \theta_i \cdot \sum_{t=1}^{T} w_{i,t}\left(x(t)\right)$ as an illustrative example, we can set a higher $\theta_i \ge 0$ if data center i is located in a severely water-stressed and drought-prone region. In line with the principle of proportionality, the total carbon footprint $\sum_{t=1}^{T} c_{i,t}(x(t))$ in $\mathcal{H}_{i,c}()$ and water footprint $\sum_{t=1}^{T} w_{i,t}(x(t))$ in $\mathcal{H}_{i,w}()$ for data center i can be normalized by the maximum processing capacity M_i to avoid overly penalizing a larger data center whose environmental footprint is inevitably larger.

Note that the two functions $\mathcal{H}_{i,c}()$ and $\mathcal{H}_{i,w}()$ are general enough and can also capture the effects of additional sustainability practices that data center operators may adopt (e.g., installing solar for carbon mitigation and restoring watersheds for local water supply [28, 29]). The term $\sum_{t=1}^{T} g_t(x(t))$ in (4a) is the total energy cost. The hyperparameters $\mu_c \ge 0$ and $\mu_w \ge 0$ indicate the relative importance weights of carbon footprint equity and water footprint equity, respectively, and can be flexibly tuned to balance the impact of carbon and water footprints. For example, by setting $\mu_c = 0$, we focus solely on the negative environmental impact of AI's water footprint. In addition, we can also include into (4a) AI's other environmental impacts such as concerns with the servers' noise pollution [12].

Importantly, the two cost terms $\max_{i\in\mathcal{N}}\left[\mathcal{H}_{i,c}\left(\sum_{t=1}^{T} c_{i,t}(x(t))\right)\right]$ and $\max_{i\in\mathcal{N}}\left[\mathcal{H}_{i,w}\left(\sum_{t=1}^{T} w_{i,t}\left(x(t)\right)\right)\right]$ improve environmental equity by penalizing the greatest environmental impacts that AI model inference creates on different regions. This is fundamentally different from the existing sustainable GLB techniques that have predominantly focused on minimizing the weighted *sum* of energy costs, carbon footprint and/or water footprint [27,33,51,52,72] and, as shown in our experiments (Section 5), can even potentially exacerbate environmental inequity by aggressively exploiting the already-disadvantaged regions.

Due to their dependency on the long-term carbon and water footprints, the two equity-related costs couple all the GLB decisions over T time slots. Thus, it is challenging to solve (4a)–(4d) and optimize the equity-aware GLB decisions in an online manner without having complete information about all the future workload arrivals, energy prices, carbon efficiency, and WUE. We leave the design of an efficient online algorithm as our future research.

4.2 Extensions

Next, we discuss how to incorporate on-site carbon-free energy and heterogeneous AI models to further enrich our GLB design.

4.2.1 On-site Carbon-free Energy

To directly cut the carbon footprint, some centers have begun to install on-site carbon-free energy, such as solar power, to partially power the workloads [27, 44]. Suppose the intermittent carbon-free energy available for data center i at time t is $z_{i,t}$. Then, the amount of electricity drawn from the grid by data center i becomes $\max\{\gamma_{i,t}e_i(x(t)) - z_{i,t}, 0\}$, where $\gamma_{i,t}$ and $e_i(x(t))$ are the PUE and AI server cluster's energy consumption, respectively. Thus, the energy cost, carbon footprint, and indirect water footprint can be calculated based on $\max\{\gamma_{i,t}e_i(x(t)) - z_{i,t}, 0\}$, while the direct water consumption remains unchanged.

4.2.2 Heterogeneous AI Models

Our current problem formulation focuses on GLB decisions by assuming a single AI model for inference and does not exploit the *performance* flexibility of AI models. In practice, for the same inference service, a set of heterogeneous AI models with distinct computing resource consumption and accuracy performance are often available via model pruning and compression [73], offering flexible energy-accuracy tradeoffs. For example, there are eight different GPT-3 model sizes, ranging from the smallest one with 125 million parameters to the largest one with 175 billion parameters [3].

Suppose that there are a set $\mathcal{K} = \{1, \cdots, K\}$ of heterogeneous AI models for our considered inference service. For time t, we can dynamically choose to run one or more AI models to serve the incoming workloads. This is also equivalent to distributing the workload $\sum_{j\in\mathcal{J}} x_{i,j}(t)$ to K heterogeneous AI models within data cener i. We denote by $y_{i,k}(t) \geq 0$ as the amount of workloads distributed to AI model k in data center i. Naturally, $y_{i,k}(t) = 0$ means that the AI model k is not chosen in data center i at time t.

When deployed in data center i, the energy consumption and server resource usage of AI model k for processing workloads $y_{i,k}(t)$ are denoted by $e_{i,k}(y_{i,k}(t))$ and $r_{i,k}(y_{i,k}(t))$, respectively. Thus, the total server energy in data center i becomes $\tilde{e}_i(y(t)) = \sum_{k\in\mathcal{K}} e_{i,k}(y_{i,k}(t))$, where $y(t) = \{y_{i,k}(t) \,|\, i \in \mathcal{N}, k \in \mathcal{K}\}$ represents the collection of decisions for workload assignment to different AI models. Similarly, with heterogeneous AI models, we can re-define the carbon footprint and water footprint as $\tilde{c}_{i,t}(y(t))$ and $\tilde{w}_{i,t}(y(t))$ by replacing $e_i(x(t))$ with $\tilde{e}_i(y(t)) = \sum_{k\in\mathcal{K}} e_{i,k}(y_{i,k}(t))$ in (2) and (3), respectively.

To optimally distribute workloads to AI models with different energy-accuracy tradeoffs, we need to consider the *cost* associated with different levels of accuracy performance, since otherwise always choosing the smallest model can generally result in the lowest energy consumption. Specifically, we refer to the cost as performance cost and denote it by $s_k(y_{i,k}(t))$, whose dependency on $y_{i,k}(t)$ can be explained by noting that the performance cost is potentially more significant when more users use the model (i.e., $y_{i,k}(t)$ is larger).

Next, by combining the energy cost and performance cost, we consider a generalized *operational* cost as follows:

$$\tilde{g}_t(y(t)) = \sum_{i\in\mathcal{N}} \sum_{k\in\mathcal{K}} \left[p_{i,t}\gamma_{i,t} \cdot e_{i,k}(y_{i,k}(t)) + \phi \cdot s_k(y_{i,k}(t))\right], \tag{5}$$

where the hyperparameter $\phi \geq 0$ converts the performance cost $s_k(y_{i,k}(t))$ to a monetary value and indicates the importance of inference performance relative to the energy cost.

Finally, we formulate the GLB problem with heterogeneous AI models as follows:

$$\min_{x(t),y(t)} \sum_{t=1}^{T} \tilde{g}_t(y(t)) + \mu_c \cdot \max_{i\in\mathcal{N}} \left[\mathcal{H}_{i,c}\left(\sum_{t=1}^{T} \tilde{c}_{i,t}(y(t))\right)\right] + \mu_w \cdot \max_{i\in\mathcal{N}} \left[\mathcal{H}_{i,w}\left(\sum_{t=1}^{T} \tilde{w}_{i,t}\left(y(t)\right)\right)\right], \tag{6a}$$

$$s.t. \quad x_{i,j}(t) = 0, \ \text{ if } B_{i,j} = 0, \quad \forall\, i \in \mathcal{N}, j \in \mathcal{J}, t = 1, \cdots, T, \tag{6b}$$

$$\sum_{x\in\mathcal{N}} x_{i,j}(t) = \lambda_{i,j}, \quad \forall\, j \in \mathcal{J}, t = 1, \cdots, T, \tag{6c}$$

$$\sum_{k\in\mathcal{K}} r_{i,k}\left(y_{i,k}(t)\right) \leq M_i, \quad \forall\, i \in \mathcal{N}, t = 1, \cdots, T, \tag{6d}$$

$$\sum_{j\in\mathcal{J}} x_{i,j}(t) = \sum_{k\in\mathcal{K}} y_{i,k}(t), \quad \forall\, j \in \mathcal{J}, t = 1, \cdots, T, \tag{6e}$$

where the objective (6a) is to minimize the generalized operational cost (including both energy and performance costs) while addressing environmental inequity, the constraint (6d) means that the total resource demand must be no more than the server cluster's capacity, and the last constraint (6e) ensures that the workload assigned to each data center is always served by one of the heterogenous AI models.

5 Experiments

In this section, we report on experiments of different GLB algorithms using trace-based simulations. Our results demonstrate that `GLB-Equity` has a great potential to effectively address AI's environmental inequity that would otherwise be potentially amplified by other GLB algorithms.

5.1 Methodology

Detailed information about data center operation, such as the number of housed servers and their configurations, hourly WUE, hourly carbon efficiency, and pricing contracts for electricity are generally considered to be business secretes and are not available in the public domain. Thus, in line with the prior GLB literature [33, 50–52], we run simulations by scaling up workload traces collected from public sources and considering synthetic data center settings that approximate realistic scenarios.

5.1.1 Workload Trace

We consider an inference service provided by homogeneous AI models (Section 4.1). To obtain the workload trace, we extract the GPU power usage data from [8] for the server cluster hosting the large language model BLOOM over an 18-day period (between September 23 and October 11 in 2022). As there is only a single workload trace provided for BLOOM in [8], we evenly distribute the workload trace to the 10 gateways (plus a small perturbation to account for different time zones), since each data center in our setup has the same capacity (Section 5.1.2) and this allows us to focus on the environmental impact without introducing unnecessary heterogeneities. As in [8], we directly quantify the amount of workload using power demand. We also scale up the workload trace to match our data center power capacity as introduced below.

5.1.2 Data Centers

We consider a set of 10 geo-distributed data centers, including four in the U.S. (Virginia, Georgia, Texas, and Nevada), four in Europe (Belgium, the Netherlands, Germany, and Denmark), and two in Asia (Singapore and Japan). These locations are all a large presence of data centers, including Google's data centers [4]. The details of data center locations are available in the appendix.

Assuming that there are 10 gateways corresponding to the 10 data center locations, we consider two scenarios: (1) **full GLB flexibility**: the workloads can be flexibly dispatched from any gateway to any data center; and (2) **partial GLB flexibility**: the workloads arriving at a gateway can only be dispatched to a certain subset of data centers. The "full GLB flexibility" scenario represents an ideal case that data center operators strive to achieve, where the "partial GLB flexibility" scenario accounts for various constraints such as network bandwidth and transmission latency.

For processing AI inference workloads, we assume that each data center houses a cluster of 500 homogeneous servers. Each server is equipped with four NVIDIA A100 GPUs and has a maximum total power of 2 kW. Thus, excluding the network switches and servers for other services beyond the scope of our study, each data center has a maximum power of 1 MW for AI inference.

We set the data center PUE as 1.1, which is consistent with the state-of-the-art PUE value with efficient operation [4,28]. For simplicity, we use the actual carbon footprint and water footprint to measure the regional environmental impact (i.e., $\mathcal{H}_{i,c}(x) = x$ and $\mathcal{H}_{i,w}(x) = x$ in (4a)).

5.1.3 Energy Price, Carbon Efficiency, and WUE

We collect hourly energy prices for the 10 data centers over the same 18-day period as our workload trace. Specifically, for each data center in Europe and Asia, we collect the hourly country-level energy prices from [74]. For the U.S. data centers, we collect the hourly energy prices from their respective ISOs [75].

For each of the U.S. data centers, we collect the state-level hourly energy fuel mix data [75] and calculate the hourly carbon efficiency and indirect WUE based on the fuel mix by following [52] and [33], respectively. The carbon efficiency and energy water intensity factor (EWIF) for each fuel mix are chosen based on [52] and [14]. We do not have free access to the hourly energy fuel mix data for our data center locations in Europe

and Asia [74]. Thus, we generate synthetic hourly fuel mixes for these locations based on the U.S. data. The details are available in the appendix.

To model the on-site WUE, we assume that the data centers use cooling towers, which are very common in the industry (even in water-stressed regions like Singapore [76] and Arizona [34]). We collect the hourly weather data from [77] for the airports closest to each of our data center locations, and then obtain the wet bulb temperature from the dry bulb temperature and relative humidity based on [78]. Next, we calculate the on-site WUE using the empirical formula in terms of the wet-bulb temperature presented in [33]. While we assume cooling towers as the heat rejection mechanism, our study can be easily generalized to air-side economizers, which use water for humidity control or when the outside dry bulb temperature is high [31].

5.1.4 Optimizer

We focus on offline optimization to quantify the potential of equity-aware GLB to address AI's environmental inequity, while leaving the design of online GLB algorithms as our future work to investigate how much of the potential can be realized with online information in practice. Specifically, we consider hourly GLB decisions and use *cvxpy* to solve (4a)–(4d) offline based on the complete information about all the future workload arrivals, energy prices, carbon efficiency, and WUE values. We refer to this offline algorithm as `GLB-Equity`. It takes about 3 minutes on a desktop with Intel i7-9700K CPU and 16GB RAM to solve the problem for an 18-day simulation in our experiments. The weight hyperparameters in (4a) are set as $\mu_c = 1500$ \$/ton and $\mu_w = 60$ \$/m^3. Note that these hyperparameters are only used to adjust the relative importance of different cost terms in the optimization process and do not reflect the true monetary costs of carbon or water footprints.

5.1.5 Metrics

We evaluate `GLB-Equity` in terms of the following metrics: **average energy cost** (the total energy cost throughout the 18-day period divided by 10 data center locations), **average carbon/water footprint** (the total carbon/water footprint throughout the 18-day period divided by 10 data center locations), and the **maximum regional carbon/water footprint** over the 18-day period among the 10 data center locations. If scaled up by a factor of 10, the average value is equivalent to the *total* value.

5.2 Baseline Algorithms

We consider the following GLB-related algorithms as baselines for comparison.

• `GLB-Cost`: This algorithm is based on [50,51,53] and only minimizes the total energy cost. It is a special case of `GLB-Equity` by setting $\mu_c = 0$ and $\mu_w = 0$ in (4a).

• `GLB-Carbon`: This algorithm only minimizes the total carbon footprint.

• `GLB-Water`: This algorithm only minimizes the total water footprint.

• `GLB-C2`: This algorithm is based on [52] and minimizes the weighted sum of the total energy cost and carbon footprint.

• `GLB-All`: This algorithm is based on [33] and minimizes the weighted sum of the total energy cost, carbon footprint, and water footprint.

• `GLB-Null`: This algorithm is a special case of GLB and directly routes workloads from each gateway to its nearest data center. It is commonly used in practice as a default baseline algorithm [52,53].

The weights for carbon and water (if applicable) in `GLB-C2` and `GLB-All` are set such that their respective total carbon and water footprints are smaller than those of `GLB-Equity`.

5.3 Results

We show our results in Table 1 by considering two different scenarios: full and partial GLB flexibilities. Our results highlight that `GLB-Equity` can improve AI's environmental equity by reducing the environmental impact on the most disadvantaged region while still keeping the average environmental footprint and energy cost close to or even lower than those of alternative GLB algorithms. Next, we discuss our results in detail.

5.3.1 Full GLB Flexibility

We first consider the full-flexibility scenario in which the workloads can be dispatched to any data center. Among all the algorithms, `GLB-Equity` has the lowest carbon and water footprints for the most disadvantaged regions. Meanwhile, the average energy cost, carbon footprint, and water footprint of `GLB-Equity` are comparable to or even lower than the other GLB algorithms. Thus, `GLB-Equity` has the lowest "maximum to

Table 1: *Comparison of GLB approaches in terms of the energy cost, carbon footprint and water footprint. The results of* `GLB-Equity` *are bolded.*

GLB Flexibility	Metric		Algorithm						
			`GLB-Cost`	`GLB-Carbon`	`GLB-Water`	`GLB-C2`	`GLB-All`	`GLB-Null`	`GLB-Equity`
Full	**Energy** (US$)	avg	29170	45535	56184	31272	31474	47038	**33669**
	Water (m^3)	avg	1525.1	1365.9	1243.9	1467.8	1429.5	1446.7	**1432.8**
		max	2607.5	2530.2	2010.4	2669.6	2566.6	2090.9	**1818.7**
	Carbon (ton)	avg	108.71	93.07	99.63	100.22	101.13	109.02	**104.91**
		max	182.14	158.44	203.56	175.56	179.17	153.37	**118.56**
Partial	**Energy** (US$)	avg	29659	45535	53976	31652	31836	47038	**34186**
	Water (m^3)	avg	1524.1	1365.9	1249.9	1464.3	1425.6	1446.7	**1432.8**
		max	2616.1	2530.2	2028.4	2675.7	2568.9	2090.9	**1860.8**
	Carbon (ton)	avg	108.32	93.07	98.46	99.78	100.76	109.02	**104.45**
		max	182.77	158.44	203.56	175.45	179.37	153.37	**117.63**

average" ratio in terms of both the carbon footprint and water footprint, reducing the regional disparity and improving environmental equity.

Interestingly, while `GLB-Cost`, `GLB-Carbon`, and `GLB-Water` can minimize the total energy cost, carbon footprint, and water footprint, respectively, they can even amplify the environmental inequity compared to `GLB-Null`. This is due to the inequity *unawareness* of these algorithms — their aggressive exploitation of certain regions may come at the cost of harming these regions in terms of environmental impacts. For example, `GLB-Cost` exploits the cheaper energy price of Texas by assigning more workloads to this region, but this can result in a disproportionately high environmental footprint in Texas due to its worse carbon efficiency and/or WUE than some other regions. While `GLB-C2` and `GLB-All` can balance the energy cost and environmental footprints in terms of the average/total metric, they can still result in disproportionately high environmental burdens on the already-disadvantaged regions due to the unawareness of equity. This is similar to algorithmic unfairness against disadvantaged individuals or user groups caused by an AI model that purely minimizes the average loss [35, 63].

While the prior studies [33, 52] have demonstrated that the total carbon footprint and water footprint are often in tension with the energy cost, our results further add that environmental equity may not be cost-free either. Nonetheless, by balancing the energy cost and environmental equity as formulated in (4a), the extra price we pay for equity can be reasonably low compared to equity-unaware `GLB-C2` and `GLB-All`.

5.3.2 Partial GLB Flexibility

Now, we consider the partial-flexibility scenario in which intra-continental workload routing is fully flexible but inter-continental workload routing is partially restricted. Specifically, we only allow partial inter-continental workload routing as follows: workloads can be flexibly routed between Asia and the western U.S. (Nevada), and between Europe and the eastern U.S. (Virginia and Georgia).

Our results are similar to those in the full-flexibility scenario. Specifically, while the inter-continental workload routing restriction limits the GLB decision space, `GLB-Equity` still has the lowest carbon and water footprints for the most disadvantaged regions. Meanwhile, the average energy cost, carbon footprint, and water footprint of `GLB-Equity` are all comparable to or even lower than the other GLB algorithms. Thus, even without full flexibility, `GLB-Equity` demonstrates a great potential to address AI'e environmental inequity in today's geographically distributed data center infrastructures.

`GLB-Null` does not route workloads across data centers and hence is not affected by the partial GLB flexibility. Interestingly, the result of `GLB-Carbon` is not affected by the inter-continental workload routing restriction in our setting, because the workloads from each continent can be processed in at least one low-carbon data center in our setup (see Table 3).

6 Concluding Remarks

In this paper, we take a first step to address the emerging environmental inequity of AI by balancing its regional negative environmental impact in an equitable manner. Concretely, we focus on the carbon and water footprints of AI model inference and propose equity-aware GLB to explicitly address the environmental impact on the most disadvantaged region. We also consider more advanced settings where there is on-site

carbon-free energy available to power the AI workloads and where we can further exploit AI's energy-accuracy flexibility by dynamically choosing one or more AI models to serve the workloads. We run trace-based simulations by considering a set of 10 geographically-distributed data centers that serve inference requests for a large language AI model. The results highlight that, compared to the existing GLB approaches, our proposed equity-aware GLB can significantly reduce the regional disparity in terms of AI's carbon and water footprints.

Our work demonstrates the need and great potential of equity-aware GLB to address AI's emerging environmental equity. An important future research problem is how to design an efficient online GLB algorithm to realize the potential in practice. This is a challenging problem that has not been well studied by the prior literature on GLB or online optimization. The key technical challenge is that reducing AI's long-term environmental impact on the most disadvantaged region (i.e., minimax in (4a)) requires all the future information, such as future AI workloads, which is only revealed sequentially in practice. Additionally, our work also opens up multiple new research directions to further improve AI's environmental equity, such as how to jointly optimize GLB and non-IT resource (e.g., batteries) management and how to leverage environmental science tools to quantify the impact of AI's carbon and water footprints.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[2] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2), feb 2022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[4] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.

[5] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, nov 2020.

[6] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[8] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. In *arXiv:2211.02001*, 2022.

[9] Jennifer Switzer, Gabriel Marcano, Ryan Kastner, and Pat Pannuto. Junkyard computing: Repurposing discarded smartphones to minimize carbon. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS 2023, page 400–412, New York, NY, USA, 2023. Association for Computing Machinery.

[10] OECD. Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. *OECD Digital Economy Papers*, (341), 2022.

[11] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. Act: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, page 784–799, New York, NY, USA, 2022. Association for Computing Machinery.

[12] Steven Gonzalez Monserrate. The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Winter 2022), January 2022. https://mit-serc.pubpub.org/pub/the-cloud-is-material.

[13] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications, 2022.

[14] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making AI less "thirsty": Uncovering and addressing the secret water footprint of ai models, 2023.

[15] José-Luis Cruz and Esteban Rossi-Hansberg. Local carbon policy. Working Paper 30027, National Bureau of Economic Research, May 2022.

[16] Mark Z. Jacobson. Enhancement of local air pollution by urban CO2 domes. *Environmental Science & Technology*, 44(7):2497–2502, 2010. PMID: 20218542.

[17] U.S. Environmental Protection Agency. About the U.S. electricity system and its impact on the environment.

[18] Paul Allen Torcellini, Nicholas Long, and Ron Judkoff. Consumptive water use for US power production. *National Renewable Energy Laboratory Technical Paper* (*TP-550-33905*), December 2003.

[19] Philipp Hacker. Sustainable AI regulation. In *Privacy Law Scholars Conference*, 2023, `https://arxiv.org/abs/2306.00292`.

[20] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022.

[21] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-MOE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, 2022.

[22] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance, 2023.

[23] Pengfei Xu, Xiaofan Zhang, Cong Hao, Yang Zhao, Yongan Zhang, Yue Wang, Chaojian Li, Zetong Guan, Deming Chen, and Yingyan Lin. AutoDNNchip: An automated DNN chip predictor and builder for both FPGAs and ASICs. In *FPGA*, 2020.

[24] Suyog Gupta and Berkin Akin. Accelerator-aware neural network design using AutoML. *arXiv preprint arXiv:2003.02838*, 2020.

[25] Anshul Gandhi, Kanad Ghose, Kartik Gopalan, Syed Rafiul Hussain, Dongyoon Lee, Yu David Liu, Zhenhua Liu, Patrick McDaniel, Shuai Mu, and Erez Zadok. Metrics for sustainability in data centers. In *HotCarbon*, 2022.

[26] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. Enabling sustainable clouds: The case for virtualizing the energy system. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '21, page 350–358, New York, NY, USA, 2021. Association for Computing Machinery.

[27] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. Carbon explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS 2023, page 118–132, New York, NY, USA, 2023. Association for Computing Machinery.

[28] Google. Environmental report, 2022, `https://sustainability.google/reports/`.

[29] Meta. Sustainability report, 2021, `https://sustainability.fb.com/`.

[30] Google. Water commitments, 2023, `https://sustainability.google/commitments/water/`.

[31] Meta. Sustainability — water, `https://sustainability.fb.com/water/`.

[32] U.S. Department of Engergy. What is environmental justice?, `https://www.energy.gov/lm/what-environmental-justice`.

[33] Mohammad A. Islam, Kishwar Ahmed, Hong Xu, Nguyen H. Tran, Gang Quan, and Shaolei Ren. Exploiting spatio-temporal diversity for water saving in geo-distributed data centers. *IEEE Transactions on Cloud Computing*, 6(3):734–746, 2018.

[34] Leila Karimi, Leeann Yacuel, Joseph Degraft-Johnson, Jamie Ashby, Michael Green, Matt Renner, Aryn Bergman, Robert Norwood, and Kerri L. Hickenbottom. Water-energy tradeoffs in data centers: A case study in hot-arid climates. *Resources, Conservation and Recycling*, 181:106194, 2022.

[35] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 66–76, New York, NY, USA, 2021. Association for Computing Machinery.

[36] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022.

[37] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524, New York, NY, USA, 2020. Association for Computing Machinery.

[38] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[39] Xueru Zhang and Mingyan Liu. *Fairness in Learning-Based Sequential Decision Algorithms: A Survey*, pages 525–555. Springer International Publishing, Cham, 2021.

[40] UNESCO. Recommendation on the ethics of artificial intelligence. In *Policy Recommendation*, 2022.

[41] Bogdana Rakova and Roel Dobbe. Algorithms as social-ecological-technological systems: an environmental justice lens on algorithmic audits. In *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency*, FAccT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[42] U.S. White House Office of Science and Technology Policy. Climate and energy implications of crypto-assets in the United States, September 2022.

[43] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. *IEEE Micro*, 42(4):37–47, jul 2022.

[44] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 38(2):1270–1280, 2023.

[45] Chuangang Ren, Di Wang, Bhuvan Urgaonkar, and Anand Sivasubramaniam. Carbon-aware energy capacity planning for datacenters. In *MASCOTS*, 2012.

[46] Xin He, Prashant Shenoy, Ramesh Sitaraman, and David Irwin. Cutting the cost of hosting online services using cloud spot markets. In *HPDC*, 2015.

[47] Stephen Lee, Rahul Urgaonkar, Ramesh Sitaraman, and Prashant Shenoy. Cost minimization using renewable cooling and thermal energy storage in CDNs. In *ICAC*, 2015.

[48] Jose Camacho, Ying Zhang, Minghua Chen, and Dah Ming Chiu. Balance your bids before your bits: The economics of geographic load-balancing. In *e-Energy*, 2014.

[49] Marco Brocanelli, Sen Li, Xiaorui Wang, and Wei Zhang. Maximizing the revenues of data centers in regulation market by coordinating with electric vehicles. *to appear in Sustainable Computing: Informatics and Systems*, 2014.

[50] L. Rao, X. Liu, L. Xie, and Wenyu Liu. Reducing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *INFOCOM*, 2010.

[51] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H. Low, and Lachlan L.H. Andrew. Greening geographical load balancing. In *SIGMETRICS*, 2011.

[52] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. It's not easy being green. *SIGCOMM Comput. Commun. Rev.*, 2012.

[53] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the electric bill for internet-scale systems. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, SIGCOMM '09, page 123–134, New York, NY, USA, 2009. Association for Computing Machinery.

[54] Kien Le, Ricardo Bianchini, Jingru Zhang, Yogesh Jaluria, Jiandong Meng, and Thu D. Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *SuperComputing*, 2011.

[55] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training, 2021.

[56] Chao Li, Amer Qouneh, and Tao Li. iSwitch: Coordinating and optimizing renewable energy powered server clusters. In *ISCA*, 2012.

[57] Shefy Manayil Kareem. Introducing critical new water data capabilities in Microsoft Cloud for Sustainability, 2023, `https://www.microsoft.com/en-us/industry/blog/sustainability/2023/03/22/introducing-critical-new-water-data-capabilities-in-microsoft-cloud-for-sustainability/`.

[58] Thomas Anderson, Adam Belay, Mosharaf Chowdhury, Asaf Cidon, and Irene Zhang. Treehouse: A case for carbon-aware datacenter software. In *HotCarbon*, 2022.

[59] Microsoft. Carbon-aware computing: Measuring and reducing the carbon footprint associated with software in execution. In *Whitepaper*, 2023.

[60] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

[61] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[62] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, jul 1996.

[63] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[64] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA, 2008. Association for Computing Machinery.

[65] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283:103238, 2020.

[66] Oyku Deniz Kose and Yanning Shen. Fast&fair: Training acceleration and bias mitigation for GNNs. *Transactions on Machine Learning Research*, 2023.

[67] Md Abu Bakar Siddik, Arman Shehabi, and Landon Marston. The environmental footprint of data centers in the United States. *Environmental Research Letters*, 16(6):064017, 2021.

[68] Mohammad. A. Islam, Shaolei Ren, Gang Quan, Muhammad Z. Shakir, and Athanasios V. Vasilakos. Water-constrained geographic load balancing in data centers. *IEEE Trans. Cloud Computing*, 2015.

[69] A. Shehabi, S. J. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, and W. Lintner. United States data center energy usage report. *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775*, 2016.

[70] The Green Grid. Water usage effectiveness (WUE): A green grid data center sustainability metric. *Whitepaper*, 2011.

[71] L. Tassiulas and S. Sarkar. Maxmin fair scheduling in wireless networks. In *Proceedings 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 763–772 vol.2, 2002.

[72] Kien Le, Ricardo Bianchini, Thu D. Nguyen, Ozlem Bilgir, and Margaret Martonosi. Capping the brown energy consumption of internet services at low cost. In *IGCC*, 2010.

[73] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR*, 2016.

[74] International Energy Agency (IEA). Data and statistics, `https://www.iea.org/data-and-statistics`.

[75] U.S. Energy Information Administration. Open data, `https://www.eia.gov/opendata/`.

[76] Singapore Public Utilities Board. Water efficiency benchmarks. `https://www.pub.gov.sg/savewater/atwork/WaterEfficiencyBenchmarks`.

[77] Iowa State University. Iowa environmental mesonet, `https://mesonet.agron.iastate.edu/`.

[78] D. Meyer and D. Thevenard. PsychroLib: A library of psychrometric functions to calculate thermodynamic properties of air. *Journal of Open Source Software*, 4(33):1137, 2019.

[79] Jordan Macknick, Robin Newmark, Garvin Heath, and KC Hallett. A review of operational water consumption and withdrawal factors for electricity generating technologies. *NREL Tech. Report: NREL/TP-6A20-50900*, 2011.

Appendix: Additional Details of the Simulation Setup

We do not have free access to the hourly energy fuel mix data for our data center locations in Europe and Asia. Thus, we generate synthetic hourly fuel mixes for these locations based on the U.S. data. We first obtain from [74] the average percentages of renewable and non-renewable energy in electricity generation between September 23 and October 11, 2022, for each data center location in Europe and Asia. Then, we scale the hourly energy fuel mix data in the U.S. to match the average percentages by mapping Texas' fuel mixes between June 1 and June 19, 2022, to Germany with non-renewable energy fuels scaled by 0.8503, Georgia's fuel mixes between June 1 and June 19, 2022, to Belgium with non-renewable energy fuels scaled by 1.5319, Georgia's fuel mixes between March 1 and March 19, 2022, to the Netherlands with non-renewable energy fuels scaled by 1.2759, Oregon's fuel mixes between July 1 and July 19, 2022, to Denmark with non-renewable energy scaled by 0.2657, Nevada's fuel mixes between March 1 and March 19, 2022, to Japan with non-renewable energy fuels scaled by 3.2374, Georgia's fuel mixes between May 1 and May 19, 2022, to Singapore with non-renewable energy fuels scaled by 4.4875. We choose different 18-day periods in order to de-correlate the European and Asian energy fuel mix traces from our actual U.S. data over the workload trace period (between September 23 and October 11, 2022).

We also show the estimated energy water intensity factor (EWIF) in m^3/MWh for common energy fuel types in the U.S. in Table 2 [14,79], and the details of our 10 data center locations in Table 3.

Table 2: *Estimated EWIF for common energy fuel types in the U.S.* [79].

Fuel Type	Coal	Nuclear	Natural Gas	Solar (PV)	Wind	Other	Hydro
EWIF (L/kWh)	1.7	2.3	1.1	0	0	1.8	68 (0, if excluded)

Table 3: *The detailed information of our data center locations. The estimated values are averaged over the 18-day period between September 23 and October 11, 2022.*

Country	**State/Province**	**City**	**Total WUE** (m^3/MWh)	**Carbon Efficiency** (ton/MWh)	**Energy Price** ($/MWh)
U.S.	Texas	Midlothian	5.7397	0.4011	64.931
U.S.	Virginia	Loudoun	5.9755	0.3741	77.793
U.S.	Georgia	Douglas	5.9001	0.4188	80.566
U.S.	Nevada	Storey	4.9306	0.2980	84.738
Germany	Hessen	Frankfurt	4.5889	0.3295	315.233
Belgium	Hainaut	Saint-Ghislain	4.9316	0.4802	247.083
Netherlands	Groningen	Eemshaven	3.0928	0.4454	248.258
Denmark	Fredericia	Fredericia	3.8900	0.1391	213.773
Japan	Chiba Prefecture	Inzai	2.4989	0.3280	129.269
Singapore	Singapore	Jurong West	5.8652	0.5260	155.462