Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud

Philipp Wiesner
wiesner@tu-berlin.de
Technische Universität Berlin
Berlin, Germany

Ilja Behnke
i.behnke@tu-berlin.de
Technische Universität Berlin
Berlin, Germany

Dominik Scheinert
dominik.scheinert@tu-berlin.de
Technische Universität Berlin
Berlin, Germany

Kordian Gontarska
kordian.gontarska@hpi.de
HPI, University of Potsdam
Potsdam, Germany

Lauritz Thamsen
lauritz.thamsen@tu-berlin.de
Technische Universität Berlin
Berlin, Germany

ABSTRACT

Depending on energy sources and demand, the carbon intensity of the public power grid fluctuates over time. Exploiting this variability is an important factor in reducing the emissions caused by data centers. However, regional differences in the availability of low-carbon energy sources make it hard to provide general best practices for when to consume electricity. Moreover, existing research in this domain focuses mostly on carbon-aware workload migration across geo-distributed data centers, or addresses demand response purely from the perspective of power grid stability and costs.

In this paper, we examine the potential impact of shifting computational workloads towards times where the energy supply is expected to be less carbon-intensive. To this end, we identify characteristics of delay-tolerant workloads and analyze the potential for temporal workload shifting in Germany, Great Britain, France, and California over the year 2020. Furthermore, we experimentally evaluate two workload shifting scenarios in a simulation to investigate the influence of time constraints, scheduling strategies, and the accuracy of carbon intensity forecasts. To accelerate research in the domain of carbon-aware computing and to support the evaluation of novel scheduling algorithms, our simulation framework and datasets are publicly available.

CCS CONCEPTS

• **Social and professional topics** → **Sustainability**; • **Software and its engineering** → *Cloud computing*.

KEYWORDS

temporal workload shifting, carbon-aware scheduling, green computing, resource management, data center

ACM Reference Format:
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 *22nd International Middleware Conference (Middleware '21), December 6–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3464298.3493399

Middleware '21, December 6–10, 2021, Virtual Event, Canada

1 INTRODUCTION

Reducing the energy demand of data centers is a major concern of research and industry alike, as it is a key driver of operational expenses and largely determines the carbon footprint of cloud computing. The extent of these efforts is most evident in the fact that data center energy consumption has grown at a much slower rate over the past decade than previously assumed [40]. This success can be attributed to technological advances such as improved processor and storage-drive efficiency on the one side, but even more importantly to the steady shift of cloud computing towards highly energy-optimized hyperscale data centers [14] that already account for roughly 50 % of all compute instances [40]. Despite all the efficiency gains, data centers worldwide consumed an estimated 205 TWh of electricity in 2018, which amounts to approximately 1 % of global energy consumption, and demand is expected to rise further in the future [40].

IT industry and public cloud providers are pushing towards reducing their impact on the climate, reinforced by a global initiative

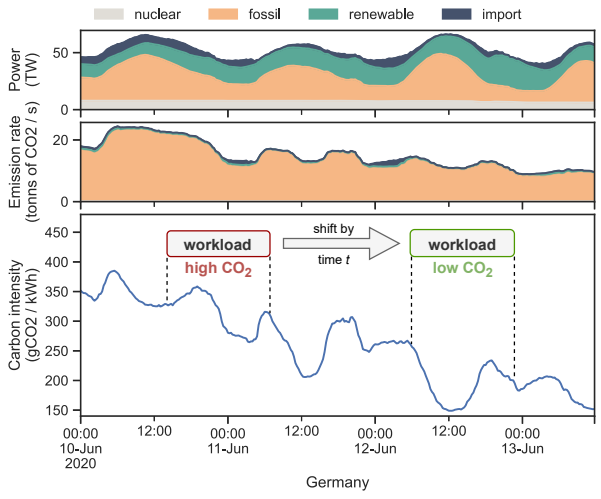

Figure 1: Power consumption, emission rate, and resulting carbon intensity in Germany, June 10-13. Scheduling workloads at times when the carbon intensity is expected to be low, can reduce the carbon footprint of data centers.

to implement carbon pricing mechanisms, such as emission trading systems (ETS) or carbon taxes [4]. However, when targeting the carbon footprint of data centers, not only the amount of energy consumed is important, but also the energy sources. For example, Google plans to operate their data centers solely on carbon-free energy by 2030 [44]. This commitment is much more extensive than what other companies tout as "carbon-free", which often only involves purchasing green power and offsetting their emissions. True carbon-free operation, on the other hand, is very hard to achieve: Given the variable nature of many renewable energy sources, such as solar and wind, operators must not only invest in energy storage systems, but also manage their demand adaptively to consume energy when and where it is emitting the least CO_2[1].

Energy sources used for electricity production vary highly in different regions, at different seasons, and at different hours of the day. This variability depends on many factors, such as weather and climate, the installed capacity of different energy sources in a region, as well as energy imports from neighboring regions. The goal of this paper is to investigate the potential impact of shifting delay-tolerant data center workloads towards times where the grid is expected to provide clean energy, as exemplary illustrated in Figure 1. To clearly state the **boundaries** of our research we note that

- the aim of this work is not to save energy but to consume energy at times, where it is generated by low-carbon sources.
- we aim at exploiting the fluctuation of carbon intensity in the public power grid and do not address the integration of local power generation that provides the data center with its own energy.
- we observe the potential of rescheduling on the time dimension. We do not consider any forms of load migration between geo-distributed data centers.

Although temporal workload shifting is already finding its way to production environments [47], existing work in the domain of carbon-aware scheduling mostly focuses on either the integration of renewable on-site or off-site installations [2, 3, 17, 20, 21, 34, 36, 36, 64, 67] or on geo-distributed load migration [41, 65, 66]. Research in the domain of data center demand management, which often utilizes load-shifting techniques, does not consider the caused carbon emissions but only addresses grid stability and energy prices [5, 13, 31, 37]. The practicability of temporal load-shifting with the goal to consume cleaner energy from the public power grid has only recently been demonstrated by Google's Carbon-Intelligent Computing System [48] (CICS). However, there does not yet exist any publicly available insights on the potential and theoretical limitations of this approach.

[1] Albeit being the most prominent source of global warming, carbon dioxide (CO_2) is not the only gas responsible for climate change. Hence, to provide a common scale for describing all greenhouse gases, a popular unit of measurement is the so called *carbon dioxide equivalent*, often abbreviated as CO_2eq. For any gas it is defined as the amount of CO_2 that would be needed to warm the earth equivalently. For simplicity, in this article we refer to CO_2eq when talking about CO_2 or carbon emissions.

Addressing this gap, we make the following **contributions**:

- we identify and categorize different characteristics of time-shiftable workloads in data centers.
- we define a methodology for estimating the regional carbon intensity of the public power grid using electricity production and inter-regional power flow data.
- we analyze the carbon-saving potential of temporal workload shifting in four regions, namely Germany, Great Britain, France, and California.
- we experimentally evaluate two scenarios via simulation, examining the impact of time constraints, scheduling strategies, and the accuracy of forecasts.
- we make all data sets and code used for the analysis and experiments of this paper publicly available[2].

The remainder of this paper is structured as follows. Section 2 discusses different characteristics of time-shiftable workloads. Section 3 explains our methodology for the following analysis and evaluation. Section 4 analyzes the theoretical potential for temporal workload shifting in four different regions. Section 5 experimentally evaluates two selected workload shifting scenarios via simulation. Section 6 reviews the related work. Section 7 concludes the paper.

2 SHIFTABLE WORKLOADS

The most important properties for determining a workload's shifting potential are its time constraints. While many workloads are expected to be finished as soon as possible, others may be subject to a degree of flexibility. However, there are further properties, such as duration, execution time, and interruptibility, can have a substantial impact on whether and how a workload can be shifted in time. This section categorizes workloads based on these characteristics. The characteristics are experimentally evaluated regarding their impact in Section 5. The terms workload and job are used interchangeably in this and the following sections.

2.1 Duration

While there is no consistent terminology, analyses of large cluster traces [24, 39, 49, 56] broadly categorize workloads into *short-running*, *long-running*, and *continuously running*.

2.1.1 Short-Running Workloads. Workloads executed in data centers are predominantly short-running. An analysis has shown, that the majority of jobs in the Google cluster traces of 2011 last only a few minutes [49]. Similar findings were made on Alibaba cluster traces, where more than 90 % of batch jobs run less than 15 minutes [39], and are more likely to be deferred or evicted due to low priority levels [24]. The shifting potential of such workloads highly depends on their time constraints. Most short-running workloads, such as Function-as-a-Service (FaaS) executions [52] or CI/CD runs [25], are expected to be finished in a timely manner. Even when delays of a few hours are tolerable, the expected potential for shifting is comparably small, as carbon intensity usually does not change quickly in large electrical grids. However, some batch jobs, such as nightly backups, may be accompanied by service-level agreements (SLAs) that allow for greater flexibility regarding the execution time. In these cases, the relative shifting potential is very high since

[2] Github: https://github.com/dos-group/lets-wait-awhile

the entire job can be moved to times of lower carbon intensity, and not only parts of it.

2.1.2 Long-Running Workloads. Analyses of Google cluster traces reveal that while only 7 % of all workloads run at production priority, a majority of these jobs are long-running [49]. Thus, the resource and memory consumption of all jobs entail a heavy-tailed distribution, where a small portion of jobs consumes most of the resources [49, 56]. Moreover, as shown on Alibaba cluster traces, long-running and prioritized workloads are likely to request significantly more resources and memory than they actually utilize [39]. For our paper, we define long-running workloads to have runtimes of up to several days. General examples for such jobs are machine learning trainings, scientific simulations, or big data analysis jobs. These workloads bear a notable absolute shifting potential since they are often very energy-intensive. Moreover, it is often humans that rely on their results to take further action. So, in practice, in many cases it makes no difference whether the issued job is finished in the middle of the night or the following morning. This flexibility can be exploited by shifting workloads without interfering with the user's workflow.

2.1.3 Continuously Running Workloads. Many computational workloads, like user-facing APIs, effectively run indefinitely by design and cannot be interrupted. Apart from these so called continuous services, there exist other computationally intensive workloads such as blockchain mining, protein folding, brute force attacks, or very long-running scientific simulations, that execute over weeks and months, or do not have any defined end date. As an example, 2000 jobs of the Google cluster traces from 2011 run for the entire trace period of 30 days [49].

Although blockchain mining in particular has received great attention for its immense power consumption [32, 35], we do not consider these workloads as shiftable in this paper, as they have no deadline or a deadline very far in the future. This paper only covers workloads up to several days, as real carbon intensity forecasts are based on weather and electricity demand forecasts which also only extend a few days into the future [7, 8, 38].

2.2 Execution Time

The expected execution time of a workload and how strictly it should be enforced are important aspects in determining its shifting potential. We therefore elaborate two categories of execution time that are illustrated in Figure 2.

2.2.1 Ad Hoc Workloads. A large number of workloads, short- and long-running, are issued in an ad hoc manner. Although some of them might follow a certain distribution which can be estimated by time series forecasting, it is not known upfront when exactly a specific job will be issued. Examples are again FaaS executions, CI/CD runs, machine learning trainings, and other jobs triggered by external events or issued by users for direct execution. The shifting potential of such workloads is limited to the future. In other words, only once a job is issued the scheduler can decide whether to execute the job immediately, or to postpone it under consideration of its time constrains.

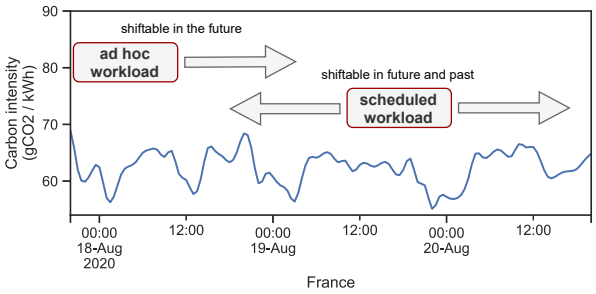

Figure 2: Scheduled workloads can potentially be shifted in both directions of time, while ad hoc workloads can only be deferred into the future.

2.2.2 Scheduled Workloads. We define scheduled workloads to be workloads that are planned to execute at a future point in time. Prominent examples are periodically scheduled batch jobs such as nightly integration test suits, nightly builds, periodic backups, updates of search indices in databases, and auto-generated reports. According to related work, a large number of jobs are recurring at fixed intervals. For example, when comparing the Google cluster traces from 2011 to the traces of 2019, it can be observed that the workload mix changed towards scheduled batch jobs while the scheduling rate increased significantly [56]. At Microsoft, periodic batch jobs have been reported to make up 60 % of processing on large clusters [27]. More than 40 % of these jobs run on a daily basis, while other frequently used periods are fifteen minutes, an hour, and twelve hours. Another study revealed that recurring jobs make up roughly 40 % of the jobs as well as cluster hours on all production clusters used for Microsoft's Bing service [1].

Scheduled workloads can, depending on their time constraints, be shifted in both directions in time. For example, a nightly job which is usually executed periodically at 1 am, could also be scheduled at a more flexible time window between 23 pm and 3 am.

2.3 Interruptibility

While certain workloads incorporate checkpoint mechanisms or store intermediate results and can thus be paused and resumed at a later point in time, other workloads must be executed without interruption. As the interruptibility of workloads can be exploited by carbon-aware schedulers as depicted in Figure 3, to better align

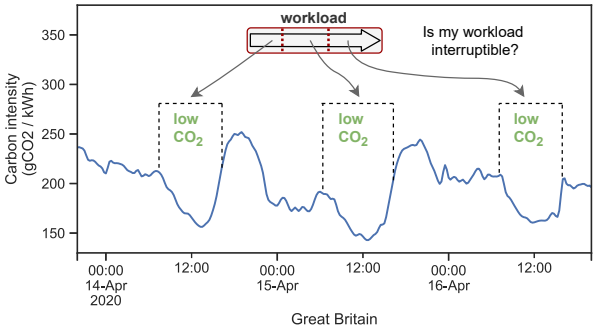

Figure 3: Interruptible workloads can be divided into chunks and scheduled separately.

Table 1: Carbon intensity of different energy sources according to [42].

Energy Source	Biopower	Solar Energy	Geothermal Energy	Hydropower	Wind Energy	Nuclear Energy	Natural Gas	Oil	Coal
gCO_2/kWh	18	46	45	4	12	16	469	840	1001

the load to times of low-carbon energy, we categorize workloads according to their interruptibility.

2.3.1 Interruptible Workloads. The possibility of pausing and resuming jobs is frequently seen in long-running workloads. Prominent examples are iterative machine learning trainings or discrete-event simulations, which often periodically write checkpoints for later analyses, resumption from earlier states, and error handling. By using such checkpoint mechanisms and state handling, it is possible to interrupt and resume workloads at a later point in time [18, 51]. Further examples of interruptible workloads include jobs that consist of many smaller tasks, like the generation of monthly business reports for different clients. As the carbon intensity of large, interconnected regions does usually not change with high frequency, it is not meaningful to split workloads in very small chunks. From this follows that the overhead, which arises when stopping and starting jobs, can often be neglected.

2.3.2 Non-Interruptible Workloads. Other workloads cannot be interrupted or interrupting them is not practical because the energy cost of starting and stopping the work outweighs the expected benefit. Examples include certain CI/CD or compile jobs that often run in freshly created, encapsulated environments which need a significant amount of time for setup and tear-down. Database migrations and backups are usually required to execute in one go to avoid data inconsistencies. Additionally, many test suits cannot be interrupted by design, for example, when they test a system under load. Non-interruptible workloads always have to be scheduled in one consecutive period and are, hence, less flexible when it comes to avoiding local maxima in carbon intensity.

3 REGIONAL CARBON INTENSITY

This section describes our methodology for selecting the analyzed regions, collecting data, and calculating the average carbon intensity of regions over time. As we want to publish all used datasets, we did not use commercially available data such as offered by services like *electricityMap*[3]. All following analyses and experiments base on the data described in this section.

3.1 Region Selection

Our analysis covers four different regions: Germany, Great Britain, France, and California. Regions were selected by the following three criteria:

(1) *Representativeness*: To represent relevant locations for data center operation, we only chose regions in which the three biggest public cloud providers - AWS, Microsoft Azure, and Google Cloud - offer regions or availability zones, or have publicly announced plans to launch operations in the near future.
(2) *Availability of data*: For our analysis we require access to each region's electricity production data by energy source with at least hourly reporting interval for the entire year 2020.
(3) *Regional diversity*: Selected regions should have different characteristics regarding types and extent of utilized energy sources as well as geographic location, to represent a diverse spectrum of regional differences.

Unfortunately, the second criteria eliminates many candidate regions because the availability of open access data on electricity production by energy source is limited. We would have liked to include regions from the southern hemisphere and emerging markets such as Brasil, South Africa, India, Korea, Japan, or Australia. However, for none of these regions it is currently possible to access historical data in the quality required for this study. Consequently, Criteria (3) is only fulfilled partially: While our selected regions do have diverse characteristics, all are located in Europe or the US.

3.2 Carbon Intensity of Energy Sources

The carbon intensity (gCO_2/kWh) of an energy source describes the amount of carbon emitted per kWh of electricity produced. There exist numerous studies on the carbon intensity of different energy sources that use slightly varying methodologies and base their estimates on different data. We base our research on carbon intensity estimates that take into account the whole life-cycle of energy sources. In particular we use the data from a comprehensive IPCC literature review that determined the median carbon intensity value stated by hundreds of different studies [42]. The values are presented in Table 1.

3.3 Carbon Intensity of Regions

To better represent the carbon emissions data centers cause by *consuming* energy, additionally to regional energy production we also

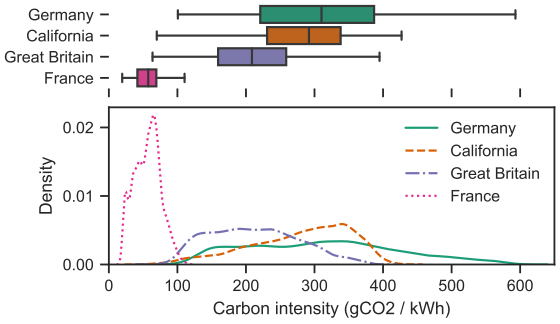

Figure 4: Distribution of carbon intensity values in the four observed regions in 2020.

[3] electricitymap.org, accessed 2021-09-21

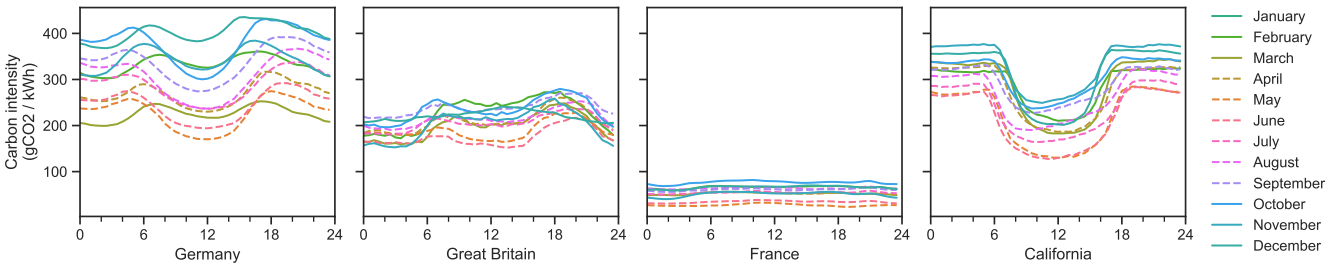

Figure 5: Daily mean carbon intensity of Germany, Great Britain, France and California by month. Since all regions are located in the northern hemisphere and therefore exhibit similar seasonal patterns, we use the same cyclic colormap to illustrate the differences between winter (solid lines) and summer months (dashed lines).

consider cross-regional flows of energy. The most precise method for this so called consumption-based accounting is to calculate the carbon intensity of all neighboring regions and to apply flow tracing in order to resemble the underlying physics of the power grid [57]. Since detailed energy generation data is not available for many regions and cross-border flows usually do not amount to a large fraction of the available power, we use a simplified method and only consider the yearly average of the neighboring regions to weight their contribution.

We define the average carbon intensity of a region C_t at time t by weighting the power generation $P_{g,t}$ of each energy source $s \in S$ by its respective carbon intensity c_s. As explained above, we additionally weight each energy import from neighboring regions $r \in R$ by the average carbon intensity of that region c_r. The resulting sum is divided by the sum of all generated and imported electricity:

$$C_t = \frac{\sum_{s=1}^{S} P_{s,t}\, c_s + \sum_{r=1}^{R} P_{r,t}\, c_r}{\sum_{s=1}^{S} P_{s,t} + \sum_{r=1}^{R} P_{r,t}}$$

For our analysis we consider the entire year 2020. The electricity production and cross-border flow data for all three European regions were retrieved via the ENTSO-E (European Network of Transmission System Operators for Electricity) API[4]. Data from the California region were retrieved via CAISO (California Independent System Operator)[5]. All data were adjusted to a common resolution of 30 minutes. For electricity production, we mapped the returned energy sources to the categories stated in Table 1. For cross-border flows, we used the yearly average carbon intensity of neighboring regions for 2020 [10].

3.4 Average vs. Marginal Carbon Intensity

Our methodology calculates the *average* carbon intensity of regions, namely their current electricity mix weighted by the carbon intensity of energy sources. A signal that captures the cause-effect relationship of load shifting even better is the *marginal* carbon intensity, which describes the carbon emissions of the energy source responsible for generating additionally requested electricity at given point in time.

[4]https://transparency.entsoe.eu, accessed 2021-09-21
[5]http://www.caiso.com, accessed 2021-09-21

Unfortunately, in practice it is very hard to identify this marginal energy source, as the decision of a power supplier to scale their production up or down is not centralized but usually incentivized via electricity prices. Additionally, this decision depends on a variety of further factors such as forecasted weather and demand as well as expected surplus or demand in neighboring regions. For this reason, there exist only probability-based methods to compute marginal carbon intensity whose results fluctuate depending on the region and time of day [33]. After reviewing marginal data provided by electricityMap, we consider marginal carbon intensity to be no practical signal for demand management at this point due to high uncertainties. This assumption is supported by Google's CICS, that also uses the average carbon intensity as an indicator for their load shifting efforts.

4 ANALYSIS OF THEORETICAL POTENTIAL

We examine the energy mix and resulting carbon intensity over time in Germany, Great Britain, France, and California throughout the year 2020. This section aims at identifying patterns in this data that can be exploited by temporal workloads shifting.

4.1 Region Analysis

In the following, the properties and peculiarities of the energy mix in the four selected regions as well as the statistical moments of their resulting carbon intensity are described. The distribution of carbon intensity values is displayed in Figure 4. The average carbon intensity throughout a day is presented in Figure 5 for each month and region.

4.1.1 Germany. Due to the wide adoption of wind (24.7 %) and solar power (8.3 %), one third of the German electricity production comes from highly variable, renewable sources. On the other hand, the remaining electricity mix is disproportionately dirty, as it is largely generated by burning lignite and black coal (22.8 %) as well as fossil gas (11.3 %). This discrepancy translates into both, the highest mean carbon intensity of 311.4 gCO_2/kWh across all observed regions, as well as highest variation of values, reaching from 100.7 to 593.1 gCO_2/kWh. The mean daily carbon intensity varies greatly over the year with a difference of up to 100 %. However, the inner-daily variance remains approximately equal regardless of the season. We observe that energy is usually the cleanest during mid day, when most solar energy is available, and around 2 am, when

electricity demand is generally low and fossil fuel power plants are throttled back.

4.1.2 Great Britain. Great Britain relies mainly on burning fossil gas (37.4 %), wind power (20.6 %) and nuclear energy (18.4 %). It has a comparably diverse energy mix and only roughly 8.7 % of the consumed energy is imported. The mean carbon intensity of 211.9 gCO_2/kWh and standard deviation are considerably lower than in Germany and stays approximately equal over the year. The inner-daily variance is higher in the winter months. Like in Germany, the carbon intensity is the cleanest during night time. However, due to the lower deployment of solar energy, carbon intensity does not drop as significantly during daylight hours.

4.1.3 France. The French energy mix comprises 69.0 % of nuclear power and 8.6 % of hydropower. Both of these energy sources are characterized by very low carbon emissions and low variability. Only a little more than 10 % of the electricity stems from variable renewable sources like wind and sun. As a result, the French power grid's carbon intensity is not only very low throughout the entire year, with a mean of 56.3 gCO_2/kWh, but also very steady. Likewise, the inner-daily variance is comparably low.

4.1.4 California. California generates 13.4 % of its total electricity from solar power - in the period between 8 am and 4 pm even 30.9 %. On the other hand, one third of the electricity comes from fossil gas and more than one quarter of the energy is imported from neighbouring states that have a comparably dirty energy mix. As a result the mean carbon intensity of 279.7 gCO_2/kWh is almost as high as in Germany, although the range of values is more comparable to Great Britain. Nevertheless, Figure 5 shows that California has very different characteristics than these regions. Because of the large amount of solar energy, the length of the low carbon intensity window during the day is strongly correlated with the number of hours of sunshine in a given month. The mean carbon intensity is generally lower in the summer months than in the winter months.

4.2 Weekly Patterns

As some non-urgent workloads can be postponed by multiple days, we first observe weekly seasonal patterns that can be exploited as shown in Figure 6. We observe that the daily carbon intensity behaves similar during workdays but has a clear drop during weekends. For example, carbon intensity of an average workday in Germany is 328.7 gCO_2/kWh, the average value during weekends is only 243.7 gCO_2/kWh, which is a decrease of 25.9 %. Likewise, we can observe decreased carbon intensity on weekends in Great Britain (20.7 %), France (22.2 %), and California (6.2 %).

This drop is caused by the decreased power demand on weekends which electricity providers respond to by mainly reducing the amount of power produced by fossil fuels. For instance, on average Germany produces 28.7 TW of energy on workdays and only 21.2 TW on weekends. The fact that electricity is greener on weekends across all observed regions suggests that shifting load to weekends is a promising approach in general. However, as stated above, it is limited to workloads with relaxed time constraints.

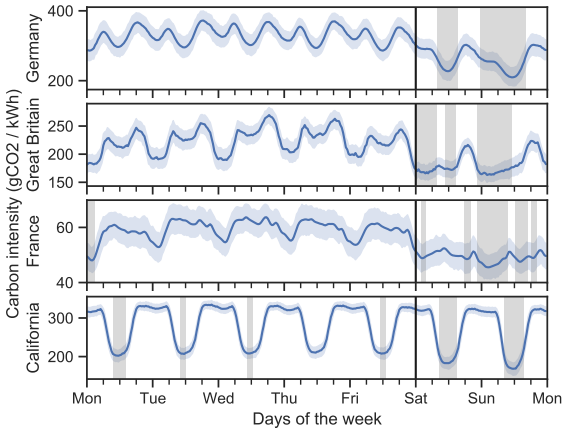

Figure 6: Mean carbon intensity during a week. The confidence interval describes the 95th percentile. Highlighted in gray are the 24 hours with lowest carbon intensity, which predominantly fall on the weekend across all regions.

4.3 Best Times of Day for Shifting

To identify the most promising times of day for shifting workloads, we define the shifting potential $p(t, W)$ at time t as follows:

$$p(t, W) = C_t - \min_{\forall t' \in W} C_{t'}$$

where W describes the forecast window, namely the set of carbon intensity data points following or preceding t. Intuitively, this function describes by how much the carbon intensity could theoretically be reduced when shifting a short-running workload at time t for up to W into the future or past. Shifting into the "past" is of course only possible for workloads that are scheduled for future execution (see Section 2.2). The carbon intensity of regions does usually not change rapidly, nor is the signal very noisy. This is why searching for the minimum value is a suitable metric here, as the chance of optimizing for negative spikes in the signal noise is very low. The presented metric only considers single data points, in other words workloads of up to 30 minutes of length, and assumes we have perfect forecast accuracy.

Figure 7 displays the shifting potential of all regions aggregated by the time of day throughout the year for four different windows: Shifting into the future and past by a maximum of two or eight hours. When considering the first row, namely shifts of up to two hours in the future, most regions exhibit little potential. An exception is California where there is a considerable shifting potential before sunrise, when carbon intensity usually drops heavily. For example, at 44 % of the days in 2020, the carbon intensity of workloads scheduled at 6 am could be reduced by more than 80 gCO_2/kWh if instead scheduled between 6 am and 8 am. Scheduled workloads can also be shifted in the opposite direction, as presented in the second row. Again, California shows the highest potential by shifting load from after to before sunset.

It becomes apparent that with bigger forecast window size the potential for improvement increases substantially. However, the optimal times for shifting differs highly in the observed regions. In Germany, we observe two times of the day that show potential

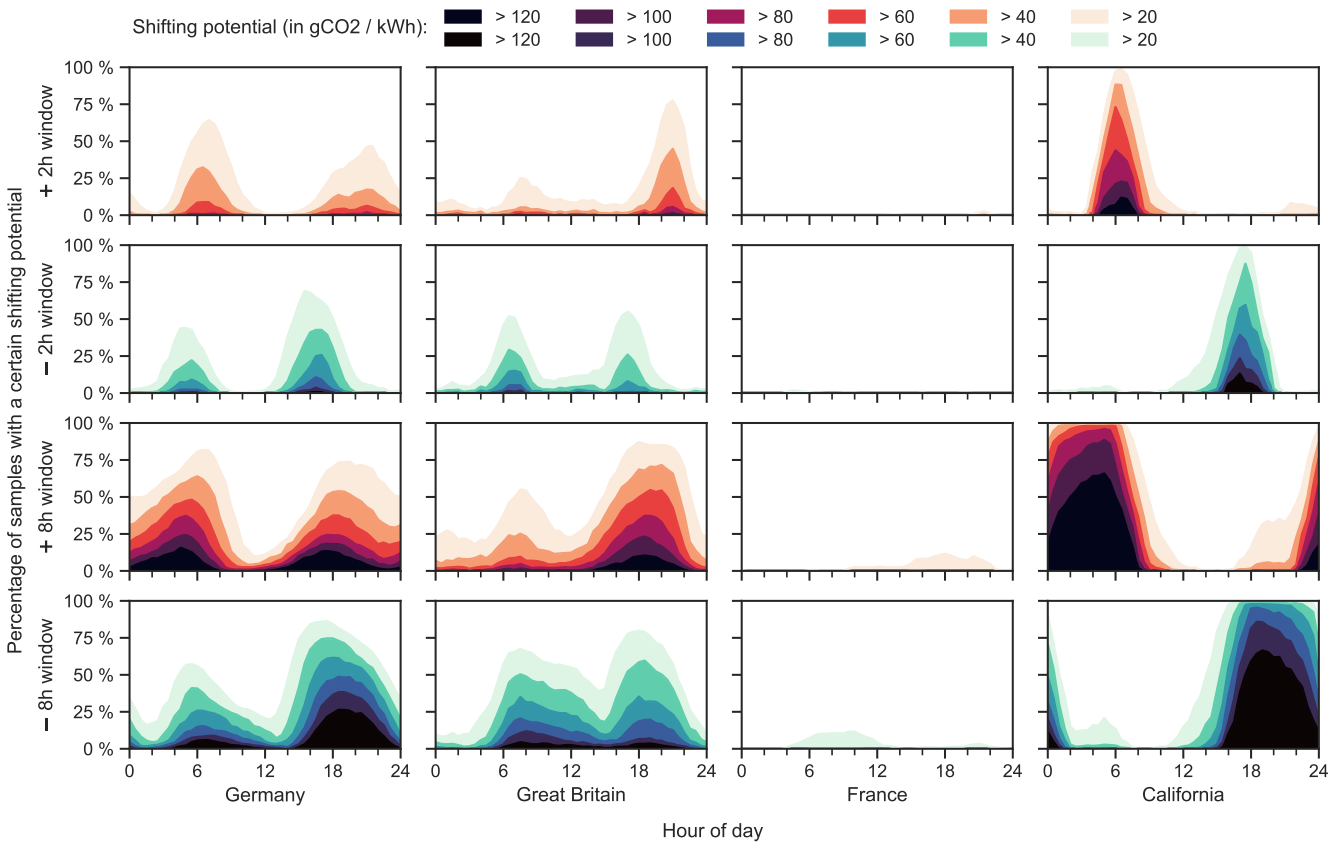

Figure 7: Potential for shifting workloads in the future (+) or past (-) at different times of the day and at two exemplary flexibility windows; 2 hours and 8 hours. For instance, the plot in column 1, row 3 describes the potential of shifting workloads in Germany by up to 8 hours into the future: The carbon intensity of 14 % of the workloads scheduled at 5 am could be reduced by at least 120 gCO_2/kWh when instead scheduled between 5 am and 1 pm.

for load shifting at 8 hour windows: In the morning hours around 7 am before sunrise and around 6 pm, escaping the high-carbon evening hours. Nevertheless, due to the high variability of energy sources in Germany, such larger forecasts offer a certain potential at virtually any time of day.

The potential for shifting workloads into the future during morning hours is considerably smaller in Great Britain, but comparably big in the evening. In general, we can observe that there is almost no potential in both directions during night time. As expected, there is barely any load shifting potential in France, even at large forecast windows. This is due to the already low carbon intensity and low variability of values during a day. In California, the potential for large forecast windows is very high during nighttime, due to the steep drop in carbon intensity during daylight hours. Consequently, workloads that are already scheduled during daytime, show little to no potential.

The key finding from this analysis is that the potential for load shifting into the future, which can be exploited by all shiftable workloads, is generally highest in the early morning hours for countries with a lot of solar power and in the evening hours for countries that throttle their fossil fuel production at night. Load shifting into the "past", which can only be exploited by future scheduled workloads, holds just as much potential and can in most cases complement load shifting into the future to attain potential throughout most parts of the day.

5 EXPERIMENTAL EVALUATION

So far, we have analyzed the theoretical potential of temporal workload shifting. In this section, we evaluate two realistic load shifting scenarios experimentally, examining the effects of time constraints, scheduling strategies, and forecast errors. Since openly available cloud computing data sets that contain information about the delay-tolerance of workloads are not available, we created two scenarios ourselves, featuring (1) short-running, periodically scheduled jobs, and (2) long-running machine learning trainings based on the StyleGAN2-ADA [28] paper. The experiments are simulated using LEAF [62], an IT infrastructure simulator that enables high-level modeling of energy consumption. The experimental setup comprises a single node, representing a data center, on which the jobs are scheduled.

5.1 Scenario I: Nightly Jobs

In the first scenario, we simulate a periodically scheduled job, such as a nightly build, integration test, or database migration. We assume these jobs to be delay-tolerant in most cases, meaning it does not make a difference to the user when exactly the job is executed, as long as it is outside of working and high-traffic hours. The aim of this experiment is to investigate the carbon saving effect of increasing the scheduling flexibility.

5.1.1 Experimental Setup. We simulate 366 periodically scheduled jobs, one for each day of the entire year 2020, with a step size of 30 minutes. Likewise, each job takes 30 minutes and is not interruptible. In the baseline experiments, jobs are scheduled to always run at 1 am. For every region, we run 16 more experiments, each increasing the time window for scheduling jobs by 30 more minutes in both directions. For example, the first shifting experiment executes all jobs between 12:30 and 1:30 am, the second between 12 and 2 am, and the last experiment schedules jobs between 5 pm and 9 am.

Since openly available, ready-to-use solutions for forecasting grid carbon intensity across different regions are not available (see Section 6.3), we added noise to the observed carbon intensity timeline in order to simulate inaccurate forecasting results. We calculated a mean absolute error of 10 for the 48-hour carbon intensity forecast by National Grid ESO [8] for 2020, which is roughly 5 % of its yearly mean. Based on this, we ran all experiments by applying normally distributed noise with $\sigma = 0.05$ times the yearly mean of the regional carbon intensity. The noise is independent of the forecast length. Since predictions in this scenario are at most 16 hours, this error can be considered an upper limit. Additionally, we repeated all experiments with optimal forecasts to investigate on the impact of errors. All experiments with forecast errors were repeated ten times and averaged.

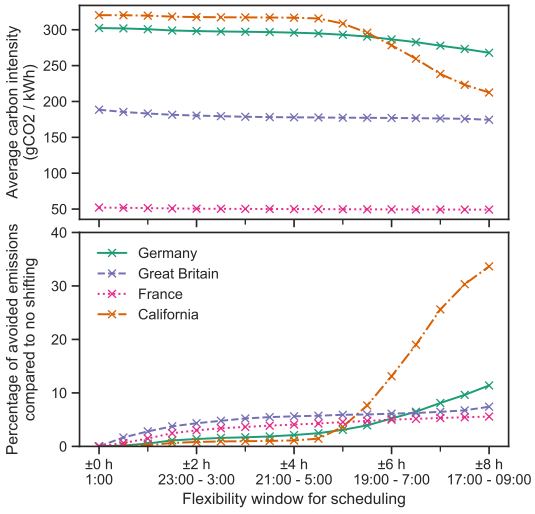

Figure 8: Average grid carbon intensity at job execution time. With increasing flexibility, the achieved carbon savings increase as well. The forecast error is 5 % in all experiments.

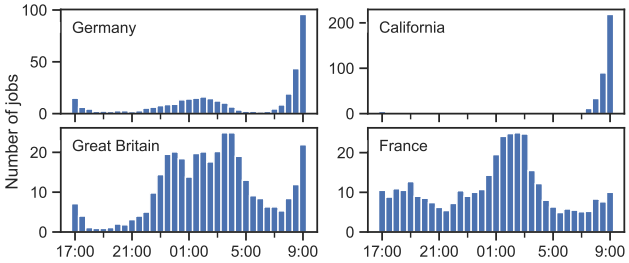

Figure 9: Number of jobs by allocated time slot for ±8 hour window size and 5 % forecast error. Germany and California shift heavily into morning hours, while Great Britain and France distribute jobs more evenly during the night.

5.1.2 Results. Figure 8 displays the experimental results. We can observe that relative to the baseline, carbon savings can be achieved across all regions. The effects differ significantly depending on the region and scheduling flexibility. For example, in France and Great Britain we can already achieve savings of 3.0 % and 4.3 %, respectively, when increasing the flexibility window by only ±2 hours. However, when the window is further enlarged, little additional savings are observed. For example, in France, the average grid carbon intensity used for powering the jobs could only be reduced by 4.1 % when considering the ±8 hour window at 5 % forecast error. In Great Britain, we managed to save 7.4 % of carbon over the year with these parameters.

Ai flexibility windows of up to ±4 hours, the resulting emissions savings for Germany and California are almost negligible. However, we observe a steep increase for windows starting at ±5 hours Even when considering forecast errors, the German scenario emits 11.2 % less carbon for the ±8 hour experiment. The forecast error has a considerable impact on this result; carbon savings were more than 2 percentage points higher with optimal forecasting. In California, the increased flexibility accounts for 13.1 % savings for the ±6 hour window and 33.7 % for the ±8 hour window under forecasts with error. The impact of errors is less significant here; optimal forecasting only improves these results by 1-1.5 percentage points.

5.1.3 Discussion. The results are consistent with our analysis on the shifting potential at different times of the day, see Figure 7. In France and Great Britain shifting potential is comparably low at night, because the mean carbon intensity at this time is already at its minimum. In contrast, in Germany and California, the potential grows significantly once the scheduler has the ability to shift workloads to the early morning or late evening hours, where they can benefit from solar energy generated during the day. This assumption is backed by Figure 9, which shows the number of jobs that were allocated to certain time slots in different regions.

For California, the case is simple: Scheduling "nightly" jobs to after sunrise significantly reduces their carbon emissions. Also in the other regions carbon-aware scheduling can reduce emissions by more than 10 %. This is not insignificant, given that the proposed shifting strategy does not have any negative impact on data center operations. From a service provider perspective, these findings can influence the design of future service-level agreements (SLAs)

and, hence, middleware systems that act within their boundraries. Providing time windows instead of fixed points in time for service execution appear to be easy-to-implement measures for reducing the carbon footprint of cloud services.

5.2 Scenario II: Machine Learning Project

The second scenario investigates the impact of different workload shifting strategies on a large machine learning project comprising a variety of different jobs. The scenario is based on the energy consumption statistics published for transparency reasons with a recent paper by NVIDIA Research introducing the StyleGAN2-ADA [28] model. The paper has received attention not only for its novelty in training generative adversarial networks, but also because the authors required 325 MWh of energy in the process of doing their research, suggesting large potential for carbon savings.

5.2.1 Experimental Setup. The authors of [28] state that 3387 machine learning jobs were executed for creating the paper, worth 145.76 GPU years. Their jobs usually run on eight GPUs, hence, an average job takes almost two days. In our scenario we assume that all jobs are scheduled ad hoc and randomly distributed across all 262 workdays of 2020 by sampling from a multinomial distribution. Each jobs is assigned a random start time during core working hours (Monday to Friday, 9 am to 5 pm). Job durations are evenly distributed between four hours and four days, resulting the same amount of GPU years as in the original project. Furthermore, we assume that job durations are known upfront accurate to 30 minutes, which is the simulation step size.

Our baseline experiment starts all jobs right when they are issued. We evaluate the potential of workload shifting in this scenario based on two time constraints:

Next Workday If jobs finish during non-working hours, they can be shifted as long as they finish before the next working day at 9 am. This allows the scheduler to take advantage of jobs that would otherwise be finished during the night or weekend without interfering with the workflow of researchers. In our scenario, this time constraint results in 20.4 % of jobs that are not shiftable because they end during working hours, 51.2 % are shiftable until the next morning and 28.4 % are shiftable over the weekend.

Semi-Weekly In practice, the individual results are often not required directly at 9 am the next day, but are evaluated in larger batches. If the time where results are actually required will be provided by users, the flexibility window for scheduling and, hence, saving potential can increase substantially. To represent this circumstance in a simple way, we assume in this time constraint that machine learning results are evaluated only twice a week. Concretely, all jobs can be shifted until the next Monday or Thursday at 9 am.

Furthermore, we want to investigate the potential benefits of exploiting incorruptible jobs, like machine learning trainings, by evaluating two scheduling strategies:

Interrupting The scheduler searches for the individual 30 minute intervals with the lowest carbon intensity and splits the job execution among these intervals.

Non-Interrupting The scheduler searches for the coherent time window with the lowest average carbon intensity and does not split the job execution.

We simulate all combinations of time constraints and scheduling strategies for each country, with a 5 % forecast error as described in Section 5.1.1.

5.2.2 Results. The carbon savings achieved by the experiments relative to the respective region's baseline experiment are depicted in Figure 10. When considering the Next Workday constraint, the Non-Interrupting scheduling managed to reduced the project's carbon emissions by 2.5 % to 6.3 %, while the Interrupting scheduling achieved reductions of 5.7 % to 8.5 %. For the Semi-Weekly constraint, Non-Interrupting scheduling saved 6.1 % to 14.4 % and Interrupting scheduling 13.3 % to 18.9 % of CO_2 emissions.

Experiments that make use of the interruptibility of machine learning jobs are improving the achieved carbon savings by 24.2 to 36.6 % for Germany, Great Britain, and France, and even by 131.2 % for California. Figure 11 shows the number of active jobs during an example period, demonstrating how Interrupting scheduling better exploits the daily fluctuation in carbon intensity than Non-Interrupting scheduling.

The additional flexibility enabled by semi-weekly scheduling causes the carbon savings to at least double across all regions,

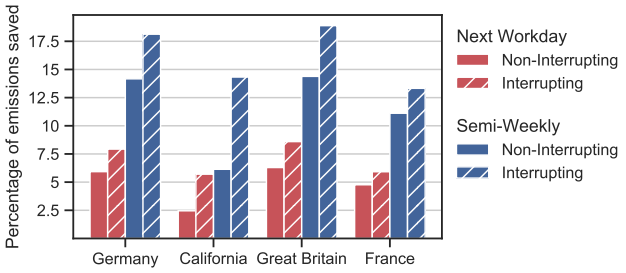

Figure 10: Carbon emission savings for different scheduling constraints and strategies by region. All experiments were simulated with 5 % forecast error.

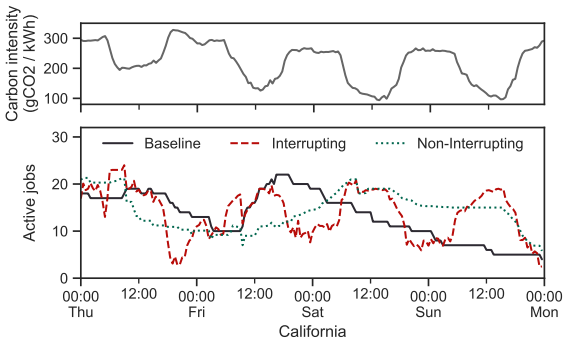

Figure 11: Number of active jobs over time for different scheduling strategies compared to the current carbon intensity. Data is from the California region, June 4-7.

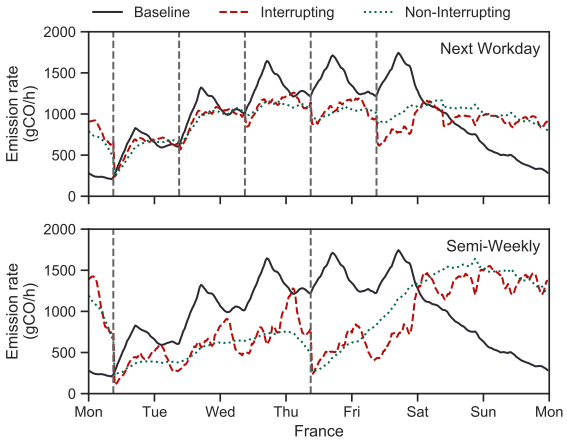

Figure 12: Average emission rates caused by different scheduling scenarios during an average week. Gray dashed lines represent the deadlines of when jobs are supposed to be finished at the two different time constraints.

compared to experiments subject to the Next Workday constraint. Figure 12 depicts how semi-weekly constraint allows the scheduler to shift even more workload towards the weekend to avoid times of high carbon intensity. Moreover, also in the Monday to Thursday period, the emission rates are significantly lower than under the Next Workday constraint.

Figure 13 displays the effect of 5 % and 10 % forecast errors on the Next Workday constraint scenario. While the savings for Non-Interrupting scheduling were almost the same independently from the applied errors, the Non-Interrupting scheduling highly benefits from low forecast errors. Findings for the Semi-Weekly scenario were equivalent.

5.2.3 Discussion. The experiments support our findings from Section 4: Shifting workloads towards nights and weekends, is a meaningful approach to consume cleaner energy. Even under time constraints that are not interfering with regular working hours, carbon savings of around 5 % are possible. With more relaxed time constraints, results improve substantially. In practice this could be implemented by letting users define the date and time until which results are actually required.

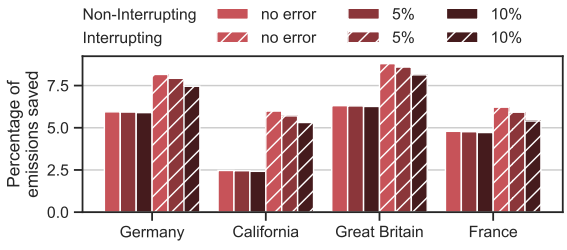

Figure 13: Influence of different forecast errors on carbon savings in the Next Workday constraint scenario.

Exploiting the interruptibility of workloads, proved to be an effective strategy to further reduce emissions. Future PaaS (Platform as a Service) and middleware systems should consider using snapshots not only for fault tolerance and possible evictions, but also for carbon-aware temporal load shifting. Forecasts with error had almost no impact on the results of the Non-Interrupting strategy but considerable impact on Interrupting scheduling. This is because Non-Interrupting scheduling optimizes for the lowest mean carbon intensity over entire intervals, and therefore is especially robust against noise in the forecasts. Interrupting scheduling is more susceptible to optimize for negative spikes, however, even with 10 % forecast errors, it always outperforms Non-Interrupting scheduling.

To conclude, we observe that experiments exploiting the interruptibility of jobs at semi-weekly scheduling are the most successful. Since, according to the data from [28], a job consumes 2036 W of power, in absolute numbers such a scheduling would have reduced the carbon emissions of the machine learning project by 8.9 t if executed in Germany and 6.3 t if executed in California or Great Britain. Although France has a very low mean carbon intensity already, savings of 1.2 t were achieved. For comparison, the per capita emissions in Germany and Great Britain in 2019 were 8.4 t and 5.5 t, respectively [50].

5.3 Limitations

In our experiments, we did not consider any resource constraints, such as the available computational capacity at a given time. While this is a reasonable assumption for Scenario I, in Scenario II there probably was a maximum number of GPUs available to the team. However, the number of active jobs in the scheduling experiments did never exceed the maximum number of active jobs of the baseline experiment by more than 42 % (64 compared to 45), which suggests that no unrealistic consolidation of workload took place.

Furthermore, forecast errors were simulated by applying uniform random noise on the actually observed carbon intensity. In reality, however, prediction errors are not uniform and also correlated. Errors grow with increasing forecast length, as well as during times with high variability such as daylight hours. Realistic forecasts may over- or underestimate the actual carbon intensity for multiple consecutive timestamps when relying, for example, on faulty weather forecasts. Because of this, the validity of our forecast error analyses are limited. A more thorough analysis applying actual forecasts in different regions would be necessary to answer important questions such as how good a forecast should be to actually request a rescheduling.

5.4 Implications

This section summarizes implications and recommendations for the future design of services, schedulers, and middleware that emerge from our evaluation.

5.4.1 Cloud and Service Providers. To exploit fluctuations in carbon intensity, providers should generally encourage users to design their workloads to be temporally flexible and/or interruptible and to declare them as such. For example, preemtive VMs (also marketed as Spot VMs/Instances) are already available across many cloud providers, offering resources at a low cost with the goal to shape load in a way beneficial to the cloud operator. As carbon

pricing mechanisms may soon account for a considerable fraction of electricity costs [4], this approach can also become profitable for carbon-aware load shaping. However, as carbon intensity characteristics and carbon pricing mechanisms vary highly from region to region, the usefulness may be limited to certain locations and has to be re-evaluated on a regular basis.

Besides direct financial incentives, service providers can also incorporate knowledge about carbon intensity patterns and the associated costs into their SLA design. For example, providing execution time windows (e.g. nightly) instead of exact times (e.g. every day at 1:00 am") for certain services increases the temporal flexibility of workloads and, hence, the carbon saving potential. Again, the data center's region plays a major role in the potential savings and has to be considered.

5.4.2 Schedulers and Middleware. Qualitative forecasts of carbon intensity and workloads are a core component of any carbon-aware scheduler. Luckily, short-term carbon intensity forecasts can often be predicted with high accuracy [7, 33]; the same applies to many data center workloads, for example, in the domain of distributed stream processing [23]. Besides, our research shows that the performance of carbon-aware schedulers highly depends on additional information about the workloads such as their temporal constraints, expected duration, and interruptibility.

Middleware will play an important role in providing this information to schedulers. On the one hand, it can offer interfaces that allow applications to conveniently declare the temporal constraints and further workload properties programatically. On the other hand, future middleware can also feature automatic detection mechanisms for certain characteristics. For instance, a middleware system could profile the time required to stop and restart a workload and automatically label it as interruptible or non-interruptible. Likewise, temporal constraints could be derived by middleware systems that, for example, know the dependency graph of tasks.

6 RELATED WORK

This section surveys related work in the field of renewable-aware workload scheduling, temporal workload shifting in the context of data center demand response, and grid carbon intensity forecasting.

6.1 Renewable-Aware Scheduling

Shaping data center load based on the availability of renewable energy has been a research topic for more than a decade [2, 22, 55, 64], with a large fraction of the literature being from the early 2010s. However, most methods focus on the integration and utilization of on-site and off-side renewable energy installations [2, 3, 6, 17, 20, 21, 34, 36, 36, 64, 67] and only few consider the carbon intensity of energy consumed from the power grid [41, 48, 65, 66]. Many approaches optimize for green energy by utilizing geo-distributed load migration, which is especially promising if data centers are being located in different hemispheres and time zones. Free Lunch [2] and GreenWare [64] are prominent examples of methods that reduce the amount of "wasted" renewable energy produced on-site by distributing workload among data centers, for example by virtual machine migration based on weather conditions. Other approaches use geo-distributed workload shifting in order to directly consume energy with lower carbon intensity [41, 65, 66].

When considering renewable-aware scheduling within single data centers, the current literature focuses on the integration of renewable energy sources and does not consider the potential reduction of carbon intensity on the public grid. For example, Aksanli et al. [3] schedule workloads by utilizing short term prediction of solar and wind energy production. GreenSlot [67] schedules batch jobs that are executed in data centers with access to on-site solar energy generation by predicting the hourly availability of solar energy two days in advance. Similarly, further approaches for renewable-aware schedulers [20, 21, 36] and works that consider the problem from a modeling [34] or user [17] perspective, do not consider the carbon intensity of the power grids, neither do related surveys and reviews [29, 43, 54].

Although Cappiello et al. already identified temporal shifting as a strategy to reduce emissions in cloud applications in 2015 [9], the first and only work utilizing this technique to date is Google's CICS [48]. They proactively shape compute load based on current and predicted power grid conditions and achieve power consumption drops of 1-2 % at times with the highest carbon intensity. However, no information on the impact of CICS in different regions is provided.

6.2 Demand Response in Data Centers

Demand response and demand-side management describe the adjustment of power usage by end-consumers during times when the power grid is stressed to capacity. The goal of demand response is to reduce peak electricity demand and, hence, to increase the stability of the power grid. In practice, this is usually achieved by providing financial incentives to consumers [15]. From an operators perspective demand response programs are therefore mainly an opportunity to reduce costs, not emissions.

Data centers have been identified as a promising industry for demand response because they consume large amounts of energy while being flexible due to their automated nature [53]. An in depth field study of data center demand response by Lawrence Berkeley National Laboratories (LBNL) [19] concludes that postponing computational load is an important demand response strategy next to load migration, shutting down or idling IT equipment, adjusting cooling, and adjusting building properties like lighting. Several works have since investigated this flexibility [5, 30, 31, 58, 61] and have proposed solutions to exploit it [11–13, 16, 31, 63]. Existing literature also considers demand response in conjunction with local power generation [37] or focuses on directly forecasting energy flexibility [59].

Data center demand response is working towards adapting the power demand profile of data centers. However, current efforts focus on power grid stability and usually optimize for cost effectiveness in incentive-based or price-based scenarios. Contrarily, our aim is to evaluate the potential of temporal workload shifting in regards to carbon savings.

6.3 Carbon Intensity Forecasts

In recent years, it has become increasingly popular to utilize carbon intensity forecasts to adaptively control power usage, for example in residential heating [45, 46, 60] or smart charging battery electric vehicles [26]. However, while there are plenty of long-term

forecasting models on CO_2 emissions of countries or industrial sectors, comparably little research exists on predicting short-term grid carbon intensity.

The most prominent supplier of carbon intensity data is Tomorrow's *electricityMap* that also provides the data for CICS. While their methodologies on real-time consumption-based carbon accounting [57] as well as short-term carbon intensity forecasting for average and marginal emissions [33] are publicly available, their data is only to a certain degree. An open carbon intensity forecast is provided by the National Grid ESO [8], a power grid operator in Great Britain. Their so called *Carbon Intensity API* provides 96 hour forecasts for different regions in Great Britain based on a rolling-window linear regression model and uses a methodology for computing carbon intensity that is similar to ours. However, their forecasting model is not open source and relies on non-publicly available weather data, meaning it cannot be transferred to other regions. Lowry [38] uses autoregressive integrated moving average (ARIMA) and neural network models for day-ahead forecasting of grid carbon intensity in order to control heating, ventilation, and air conditioning systems. Lastly, Bodke et al. [7] use a decomposition approach and forecast the grid carbon intensity of regions within Europe via statistical methods.

7 CONCLUSION

This paper examines the potential of temporally shifting computational workloads in data centers with the goal to consume cleaner energy from the public power grid. We provide an overview on characteristics of shiftable workloads and analyze the regional carbon intensity of Germany, California, Great Britain, and France over the year 2020. Our findings suggest that short-term shifting potential is often high before sunrise in countries with a lot of solar power, and in the evening hours, because most countries throttle their fossil fuel power stations at night. Moreover, shifting delay-tolerant workloads towards weekends can result in more than 20 % savings in most regions. The experimental evaluation supports our analytical findings and demonstrates that the highest savings can be achieved when relaxing time constraints and actively exploiting the interruptibility of workloads during scheduling. For example, shifting workloads whose results are not needed by the next working day can already reduce emissions by over 5 % across all regions.

Future work will address the development and evaluation of schedulers that take advantage of the findings in this paper. To this end, we hope that our simulator and published datasets will prove to be useful tools for exploring new approaches in this domain. In particular, we want to use them to research on the combination of temporal and geo-distributed scheduling, which has received little attention to date.

ACKNOWLEDGMENTS

We would like to thank all Middleware reviewers for their valuable comments and suggestions. We also express our sincere thanks to Martin Schellenberger for his insights that helped to shape this work. This research was supported by the German Academic Exchange Service (DAAD) as ide3a and the German Ministry for Education and Research (BMBF) as BIFOLD (research grant 01IS18025A).

REFERENCES

[1] Sameer Agarwal, Srikanth Kandula, Nicolas Bruno, Ming-Chuan Wu, Ion Stoica, and Jingren Zhou. 2012. Reoptimizing Data Parallel Computing. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*.

[2] Sherif Akoush, Ripduman Sohan, Andrew Rice, Andrew W. Moore, and Andy Hopper. 2011. Free Lunch: Exploiting Renewable Energy for Computing. In *13th Workshop on Hot Topics in Operating Systems (HotOS XIII)*.

[3] Baris Aksanli, Jagannathan Venkatesh, Liuyi Zhang, and Tajana Rosing. 2011. Utilizing Green Energy Prediction to Schedule Mixed Batch and Service Jobs in Data Centers. In *Proceedings of the 4th Workshop on Power-Aware Computing and Systems (HotPower '11)*.

[4] World Bank. 2020. *State and Trends of Carbon Pricing 2020.* Technical Report. Washington, DC: World Bank.

[5] Robert Basmadjian. 2019. Flexibility-Based Energy and Demand Management in Data Centers: A Case Study for Cloud Computing. *Energies* 12, 17 (2019).

[6] Nicolas Beldiceanu, Bárbara Dumas Feris, P. Gravey, Sabbir Hasan, Claude Jard, Thomas Ledoux, Yunbo Li, Didier Lime, Gilles Madi Wamba, Jean-Marc Menaud, Pascal Morel, Michel Morvan, Marie-Laure Moulinard, Anne-Cécile Orgerie, Jean-Louis Pazat, Olivier Roux, and Ammar Sharaiha. 2017. Towards energy-proportional Clouds partially powered by renewable energy. *Computing* 99 (2017).

[7] Neeraj Bokde, Bo Tranberg, and Gorm Andresen. 2021. Short-term CO2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. *Applied Energy* 281 (2021).

[8] Alasdair R. W. Bruce, Lyndon Ruff, James Kelloway, Fraser MacMillan, and Alex Rogers. 2021. *Carbon Intensity Forecast Methodology.* Technical Report. National Grid ESO.

[9] Cinzia Cappiello, Nguyen Thi Thao Ho, Barbara Pernici, Pierluigi Plebani, and Monica Vitali. 2016. CO2-Aware Adaptation Strategies for Cloud Applications. *IEEE Transactions on Cloud Computing* 4, 2 (2016).

[10] Carbon Footprint Ltd 2020. *Country Specific Electricity Grid Greenhouse Gas Emission Factors v1.4.* Carbon Footprint Ltd.

[11] Hao Chen, Michael C. Caramanis, and Ayse K. Coskun. 2014. The data center as a grid load stabilizer. In *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*.

[12] T. Cioara, I. Anghel, Massimo Bertoncini, I. Salomie, D. Arnone, M. Mammina, T. H. Velivassaki, and M. Antal. 2018. Optimized flexibility management enacting Data Centres participation in Smart Demand Response programs. *Future Gener. Comput. Syst.* 78 (2018).

[13] Tudor Cioara, Ionut Anghel, Ioan Salomie, Marcel Antal, Claudia Pop, Massimo Bertoncini, Diego Arnone, and Florin Pop. 2019. Exploiting data centres energy flexibility in smart cities: Business scenarios. *Information Sciences* 476 (2019).

[14] Cisco. 2018. *Cisco Global Cloud Index: Forecast and methodology (2016–2021).* Technical Report. Cisco.

[15] European Commission. 2013. *Incorporing demand side flexibility, in particular demand response, in electricity markets.* SWD (2013) 442 final.

[16] Lisette Cupelli, Thomas Schütz, Pooyan Jahangiri, Marcus Fuchs, Antonello Monti, and Dirk Müller. 2018. Data Center Control Strategy for Participation in Demand Response Programs. *IEEE Transactions on Industrial Informatics* 14, 11 (2018).

[17] C. Dupont, M. Sheikhalishahi, F. M. Facca, and F. Hermenier. 2015. An Energy Aware Application Controller for Optimizing Renewable Energy Consumption in Data Centres. In *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*.

[18] Kurt B. Ferreira, Rolf Riesen, Patrick G. Bridges, Dorian C. Arnold, and Ron Brightwell. 2014. Accelerating incremental checkpointing for extreme-scale computing. *Future Gener. Comput. Syst.* 30 (2014).

[19] Girish Ghatikar, Venkata Ganti, Nance Matson, and Mary Ann Piette. 2012. *Demand Response Opportunities and Enabling Technologies for Data Centers: Findings From Field Studies.* Technical Report. PG&E/SDG&E/CEC/LBNL.

[20] Íñigo Goiri, William Katsak, Kien Le, Thu D. Nguyen, and Ricardo Bianchini. 2013. Parasol and GreenSwitch: Managing Datacenters Powered by Renewable Energy. *SIGPLAN Not.* 48, 4 (2013).

[21] Íñigo Goiri, Kien Le, Thu D. Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. 2012. GreenHadoop: Leveraging Green Energy in Data-Processing Frameworks. In *Proceedings of the 7th ACM European Conference on Computer Systems (EuroSys '12)*.

[22] Í. Goiri, R. Beauchea, K. Le, T. D. Nguyen, M. E. Haque, J. Guitart, J. Torres, and R. Bianchini. 2011. GreenSlot: Scheduling energy consumption in green datacenters. In *SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*.

[23] Kain Kordian Gontarska, Morgan Geldenhuys, Dominik Scheinert, Philipp Wiesner, Andreas Polze, and Lauritz Thamsen. 2021. Evaluation of Load Prediction Techniques for Distributed Stream Processing. In *9th IEEE International Conference on Cloud Engineering*.

[24] Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. 2019. Who limits the resource efficiency of my datacenter: an analysis of Alibaba datacenter traces. In *Proceedings of the International Symposium on Quality of Service, IWQoS 2019, Phoenix, AZ, USA, June 24-25, 2019.*

[25] Michael Hilton, Timothy Tunnell, Kai Huang, Darko Marinov, and Danny Dig. 2016. Usage, costs, and benefits of continuous integration in open-source projects. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ASE 2016, Singapore, September 3-7, 2016.*

[26] Julian Huber, Kai Lohmann, Marc Schmidt, and Christof Weinhardt. 2020. Carbon efficient Smart Charging using Forecasts of Marginal Emission Factors. *Journal of Cleaner Production* 284 (2020).

[27] Sangeetha Abdu Jyothi, Carlo Curino, Ishai Menache, Shravan Matthur Narayanamurthy, Alexey Tumanov, Jonathan Yaniv, Ruslan Mavlyutov, Iñigo Goiri, Subru Krishnan, Janardhan Kulkarni, and Sriram Rao. 2016. Morpheus: Towards Automated SLOs for Enterprise Clusters. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).*

[29] Atefeh Khosravi and Rajkumar Buyya. 2017. *Advancing Cloud Database Systems and Capacity Planning With Dynamic Applications.* IGI Global, Chapter Energy and Carbon Footprint-Aware Management of Geo-Distributed Cloud Data Centers: A Taxonomy, State of the Art, and Future Directions, 27 – 46.

[30] Sonja Klingert. 2018. Mapping Data Centre Business Types with Power Management Strategies to Identify Demand Response Candidates. In *Proceedings of the Ninth International Conference on Future Energy Systems (e-Energy '18).*

[31] Sonja Klingert and Sebastian Szilvas. 2020. Spinning gold from straw - evaluating the flexibility of data centres on power markets. *Energy Informatics* 3 (2020).

[32] Max J Krause and Thabet Tolaymat. 2018. Quantification of energy and carbon costs for mining cryptocurrencies. *Nature Sustainability* 1, 11 (2018).

[33] Kenneth Leerbeck, Peder Bacher, Rune Grønborg Junker, Goran Goranović, Olivier Corradi, Razgar Ebrahimy, Anna Tveit, and Henrik Madsen. 2020. Short-term forecasting of CO2 emission intensity in power grids by machine learning. *Applied Energy* 277 (2020).

[34] C. Li, A. Qouneh, and T. Li. 2012. iSwitch: Coordinating and optimizing renewable energy powered server clusters. In *2012 39th Annual International Symposium on Computer Architecture (ISCA).*

[35] Jingming Li, Nianping Li, Jinqing Peng, Haijiao Cui, and Zhibin Wu. 2019. Energy consumption of cryptocurrency mining: A study of electricity consumption in mining cryptocurrencies. *Energy* 168, C (2019).

[36] Zhenhua Liu, Yuan Chen, Cullen Bash, Adam Wierman, Daniel Gmach, Zhikui Wang, Manish Marwah, and Chris Hyser. 2012. Renewable and Cooling Aware Workload Management for Sustainable Data Centers. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '12).*

[37] Zhenhua Liu, Adam Wierman, Yuan Chen, Benjamin Razon, and Niangjun Chen. 2013. Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems.*

[38] Gordon Lowry. 2018. Day-ahead forecasting of grid carbon intensity in support of heating, ventilation and air-conditioning plant demand response decision-making to reduce carbon emissions. *Building Services Engineering Research and Technology* 39 (2018).

[39] Chengzhi Lu, Kejiang Ye, Guoyao Xu, Cheng-Zhong Xu, and Tongxin Bai. 2017. Imbalance in the cloud: An analysis on Alibaba cluster trace. In *2017 IEEE International Conference on Big Data.*

[40] E. Masanet, A. Shehabi, Nuoa Lei, S. Smith, and J. Koomey. 2020. Recalibrating global data center energy-use estimates. *Science* 367 (2020).

[41] A. Moghaddam, Reza Farrahi Moghaddam, and Mohamed Cheriet. 2014. Carbon-aware distributed cloud: Multi-level grouping genetic algorithm. *Cluster Comput* (2014).

[42] William Moomaw, Peter Burgherr, Garvin Heath, Manfred Lenzen, John Nyboer, and Aviel Verbruggen. 2011. Annex II: Methodology. In *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation.*

[43] Eduard Oró, Victor Depoorter, Albert Garcia, and Jaume Salom. 2015. Energy efficiency and renewable energy integration in data centres. Strategies and modelling review. *Renewable and Sustainable Energy Reviews* 42 (2015).

[44] Sundar Pichai. 2020. Our third decade of climate action: Realizing a carbon-free future. *Google* (2020). https://blog.google/outreach-initiatives/sustainability/our-third-decade-climate-action-realizing-carbon-free-future (accessed 2021-05-21).

[45] Thibault Péan, Ramon Costa-Castelló, and Jaume Salom. 2019. Price and carbon-based energy flexibility of residential heating and cooling loads using model predictive control. *Sustainable Cities and Society* 50 (2019).

[46] Thibault Péan, Jaume Salom, and Joana Ortiz. 2018. Environmental and Economic Impact of Demand Response Strategies for Energy Flexible Buildings. In *4th Building Simulation and Optimization Conference (BSO 2018).*

[47] Ana Radovanovic. 2020. Our data centers now work harder when the sun shines and wind blows. *Google* (2020). https://blog.google/inside-google/infrastructure/data-centers-work-harder-sun-shines-wind-blows (accessed 2021-05-21).

[48] Ana Radovanovic, 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. 2021. Carbon-Aware Computing for Datacenters. *arXiv:2106.11750 [cs.DC]* (2021).

[49] Charles Reiss, Alexey Tumanov, Gregory R. Ganger, Randy H. Katz, and Michael A. Kozuch. 2012. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *ACM Symposium on Cloud Computing, SOCC '12, San Jose, CA, USA, October 14-17, 2012.*

[50] Hannah Ritchie and Max Roser. 2020. CO2 and Greenhouse Gas Emissions. *Our World in Data* (2020). https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions (accessed 2021-05-21).

[51] Elvis Rojas, Albert Njoroge Kahira, Esteban Meneses, Leonardo Bautista-Gomez, and Rosa M. Badia. 2020. A Study of Checkpointing in Large Scale Training of Deep Neural Networks. *CoRR* abs/2012.00825 (2020).

[52] Mohammad Shahrad, Rodrigo Fonseca, Iñigo Goiri, Gohar Chaudhry, Paul Batum, Jason Cooke, Eduardo Laureano, Colby Tresness, Mark Russinovich, and Ricardo Bianchini. 2020. Serverless in the Wild: Characterizing and Optimizing the Serverless Workload at a Large Cloud Provider. In *2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020.*

[53] M. H. Shoreh, P. Siano, M. Shafie-khah, V. Loia, and J. Catalão. 2016. A survey of industrial applications of Demand Response. *Electric Power Systems Research* 141 (2016).

[54] Junaid Shuja, Abdullah Gani, Shahaboddin Shamshirband, Raja Wasim Ahmad, and Kashif Bilal. 2016. Sustainable Cloud Data Centers: A survey of enabling techniques and technologies. *Renewable and Sustainable Energy Reviews* 62 (2016).

[55] Christopher Stewart and Kai Shen. 2009. Some Joules Are More Precious Than Others: Managing Renewable Energy in the Datacenter. In *Proceedings of the workshop on power aware computing and systems.*

[56] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E. Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. 2020. Borg: the next generation. In *EuroSys '20: Fifteenth EuroSys Conference 2020, Heraklion, Greece, April 27-30, 2020.*

[57] Bo Tranberg, Olivier Corradi, Bruno Lajoie, Thomas Gibon, Iain Staffell, and Gorm Bruun Andresen. 2019. Real-time carbon accounting method for the European electricity markets. *Energy Strategy Reviews* 26 (2019).

[58] Thiago Vasques, Pedro Moura, and Anibal Almeida. 2019. A review on energy efficiency and demand response with focus on small and medium data centers. *Energy Efficiency* 12 (2019).

[59] Andreea Valeria Vesa, Tudor Cioara, Ionut Anghel, Marcel Antal, Claudia Pop, Bogdan Iancu, Ioan Salomie, and Vasile Teodor Dadarlat. 2020. Energy Flexibility Prediction for Data Center Engagement in Demand Response Programs. *Sustainability* 12, 4 (2020).

[60] P.J.C. Vogler-Finck, R. Wisniewski, and P. Popovski. 2018. Reducing the carbon footprint of house heating through model predictive control – A simulation study in Danish conditions. *Sustainable Cities and Society* 42 (2018).

[61] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad. 2014. Opportunities and challenges for data center demand response. In *International Green Computing Conference.*

[62] Philipp Wiesner and Lauritz Thamsen. 2021. LEAF: Simulating Large Energy-Aware Fog Computing Environments. In *2021 IEEE 5th International Conference on Fog and Edge Computing (ICFEC).*

[63] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely. 2012. Data centers power reduction: A two time scale approach for delay tolerant workloads. In *2012 Proceedings IEEE INFOCOM.*

[64] Yanwei Zhang, Yefu Wang, and Xiaorui Wang. 2011. GreenWare: Greening Cloud-Scale Data Centers to Maximize the Use of Renewable Energy. In *Middleware 2011.*

[65] Jiajia Zheng, A. Chien, and S. Suh. 2020. Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers. *Joule* 4 (2020).

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

[67] Íñigo Goiri, Md E. Haque, Kien Le, Ryan Beauchea, Thu D. Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. 2015. Matching renewable energy supply and demand in green datacenters. *Ad Hoc Networks* 25 (2015). New Research Challenges in Mobile, Opportunistic and Delay-Tolerant Networks Energy-Aware Data Centers: Architecture, Infrastructure, and Communication.