On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud

Thanathorn Sukprasert
University of Massachusetts Amherst

Abel Souza
University of Massachusetts Amherst

Noman Bashir
Massachusetts Institute of Technology

David Irwin
University of Massachusetts Amherst

Prashant Shenoy
University of Massachusetts Amherst

Abstract

Cloud platforms have been focusing on reducing their carbon emissions by shifting workloads across time and locations to when and where low-carbon energy is available. Despite the prominence of this idea, prior work has only quantified the potential of spatiotemporal workload shifting in narrow settings, i.e., for specific workloads in select regions. In particular, there has been limited work on quantifying an upper bound on the ideal and practical benefits of carbon-aware spatiotemporal workload shifting for a wide range of cloud workloads. To address the problem, we conduct a detailed data-driven analysis to understand the benefits and limitations of carbon-aware spatiotemporal scheduling for cloud workloads. We utilize carbon intensity data from 123 regions, encompassing most major cloud sites, to analyze two broad classes of workloads—batch and interactive—and their various characteristics, e.g., job duration, deadlines, and SLOs. Our findings show that while spatiotemporal workload shifting can reduce workloads' carbon emissions, the practical upper bounds of these carbon reductions are currently limited and far from ideal. We also show that simple scheduling policies often yield most of these reductions, with more sophisticated techniques yielding little additional benefit. Notably, we also find that the benefit of carbon-aware workload scheduling relative to carbon-agnostic scheduling will decrease as the energy supply becomes "greener."

1 Objectives and Methodology

The primary goal of our analysis is to quantify an upper bound on carbon reduction from spatiotemporal workload shifting under ideal and constrained conditions. Our hypothesis is that while the upper bound of spatiotemporal workload shifting exhibits significant reductions in computing's carbon emissions, there exists a substantial gap between the ideal and constrained conditions. To quantify carbon reduction and evaluate our hypothesis, we focus on answering the specific research questions below. We then outline our methodology for answering these questions.

1. **Global Carbon Analysis.** What are the characteristics of grid energy's carbon-intensity worldwide? How do its magnitude, variance, and periodicity vary across regions? How has it changed in recent years?
2. **Spatial Migration:** How much carbon reduction is possible from spatially migrating workloads? How might capacity, latency SLOs, and regional privacy constraints impact this carbon reduction What is the optimal policy for minimizing carbon emissions?
3. **Temporal Shifting.** How much carbon reduction is possible from temporally shifting delay-tolerant batch workloads? How does this carbon reduction vary with workload characteristics, such as job length and slack?
4. **What-If Scenarios.** What are the benefits of combining spatial and temporal shifting, and how much carbon reduction accrues from each method? How does the i) ratio of migratable workload, ii) prediction error, and iii) increase in renewables impact the carbon reductions from temporal and spatial shifting?

1.1 Analysis Setup

Below, we provide details on our i) carbon-intensity data sources and ii) metric for quantifying carbon reduction.

1.2 Carbon-intensity Data

We collected carbon-intensity traces for 123 different geographical regions worldwide from 2020 to 2022 using the Electricity Maps web API. Each trace reports energy's average carbon-intensity, measured in grams of carbon dioxide equivalent per kilowatt-hour ($g{\cdot}CO_2eq/kWh$), in hourly granularity. The 123 locations includes our entire carbon trace dataset and encompasses 99 known datacenter locations:35 for Google Cloud Platform (GCP), 24 for Microsoft Azure, 23 for Amazon Web Services (AWS), 7 for IBM, and 10 for Alibaba.

1.3 Metrics.

We quantify carbon reduction in terms of *absolute carbon reduction* and *global average carbon reduction*. Below, we define how both metrics are calculated.

a) **Absolute Carbon Reduction** is the difference between carbon emissions after any spatiotemporal workload shifting and the carbon-agnostic baseline. We measure it in $g{\cdot}CO_2eq$, where a higher value is better.
b) **Global Average Reduction** is the average absolute carbon reduction of 123 regions from spatiotemporal workload shifting compared with the global average carbon-intensity of 368.39 $g{\cdot}CO_2eq/kWh$, expressed as a percentage.