Poster: Spatiotemporal Carbon-Aware Scheduling in the Cloud: Limits and Benefits

Thanathorn Sukprasert, Abel Souza, Noman Bashir, David Irwin, Prashant Shenoy
University of Massachusetts Amherst

ABSTRACT

As the demand for computing continues to grow exponentially and datacenters are already highly optimized, many have suggested leveraging computing workload's spatiotemporal flexibility. However, different workloads may have different degrees of flexibility, including execution deadlines, data protection laws, or latency requirements. These constraints, along with many others, limit the potential benefits of carbon-aware spatiotemporal workload shifting; the *achievable* benefits of these approaches are unclear—an aspect not addressed by prior research. Accurately quantifying the achievable benefits of carbon-aware spatiotemporal workload scheduling is critically important, as many in research and industry are already devoting significant time and resources to realize these benefits. To address the problem, we conduct a large-scale longitudinal analysis of carbon-aware spatiotemporal workload shifting to answer the following research question: *What are the maximum carbon emission reductions that can be achieved due to temporal and spatial workload shifting for different types of cloud workloads and in different parts of the world?*

CCS CONCEPTS

• **Computer systems organization → Cloud computing**.

KEYWORDS

Sustainable Computing; Carbon-Aware Systems, Scheduling

1 INTRODUCTION

The demand for computing continues to grow exponentially with the growing popularity of computationally intensive AI workloads, such as ChatGPT [2]. Since computation requires energy, the exponential growth in computing demand translates to an accelerated increase in energy consumption. Estimates indicate that cloud datacenters currently consume ~3% of global electricity and this percentage is expected to rise to 13% by 2030 [1]. Because of this, there is a growing concern that growing computation demand would subsequently lead to an increase in global carbon emissions. Unfortunately, efforts to improve the energy efficiency of cloud platforms are unlikely to have a significant impact on reducing carbon emissions because cloud platforms are already highly optimized, for datacenters currently have PUE values close to the optimal 1. As another approach to reducing carbon emissions, many have suggested leveraging computing workloads' temporal and spatial flexibility by dynamically shifting the time and location of execution to when and where jobs can be performed using low-carbon energy.

To address the problem, we conduct a large-scale longitudinal analysis of historical carbon intensity data globally to quantify the upper bound of carbon reduction through temporal and spatial workload shifting. Our analysis includes traces of hourly average carbon intensity from 123 regions worldwide and covers a three-year period from 2020 to 2022. *The primary finding from our analysis is that the upper bound on carbon savings from spatiotemporal workload shifting is often modest and that simple policies can yield most of the benefits.* Thus, contrary to conventional wisdom, i) carbon-aware spatial and temporal workload shifting is likely not a panacea for significantly reducing cloud platforms' carbon emissions, and ii) pursuing further research on sophisticated policies is likely to yield diminishing returns. We highlight many of the insights from our analysis below.

- In a real-world setting, the average carbon savings from temporal shifting are limited even if the workload has significant performance and temporal flexibility. This is because large jobs benefit the least from temporal shifting but typically dominate the resource and energy consumption in real-world workloads.
- Spatial migration offers substantial carbon savings. However, in real-world setting, the savings may be constrained by capacity limitations in the destination regions, migration overhead, latency SLOs, and privacy regulations.
- There is little need for sophisticated region-hopping policies, and migrating to the greenest available region once would yield most of the potential savings.

2 OBJECTIVES AND METHODOLOGY

In this work, our hypothesis is that spatiotemporal workload scheduling could yield significant reductions in computing's carbon emissions. To quantify and evaluate the hypothesis, we focus on answering the following research questions:

(1) **Global Carbon Analysis.** How does carbon intensity look like for different regions around the world in terms of their magnitude and variability?
(2) **Temporal Workload Shifting.** How much is the possible savings from temporally shifting delay-tolerant batch workloads? How do the savings vary with workload characteristics, such as job length and slack?
(3) **Spatial Workload Migration.** How is the possible savings from spatially migrating workloads? Do we need a sophisticated migration policy?

Our experiments explore spatiotemporal characteristics involving moving jobs across space and time to align with low-carbon energy availability. The job length indicates the time required for completion without interruptions, covering a range from interactive jobs (< 1 minute) to small batch jobs (up to a day) and long

e-Energy '23 Companion, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0227-3/23/06…$15.00
https://doi.org/10.1145/3599733.3604772

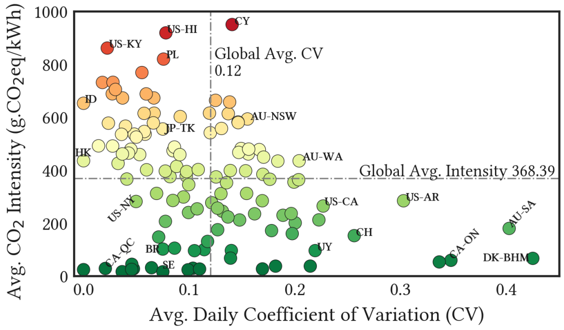
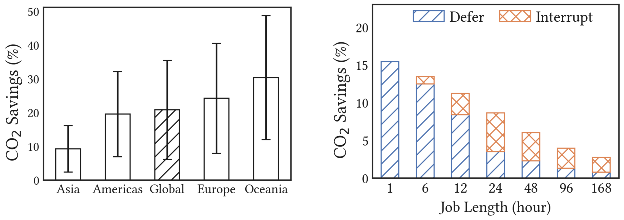


Figure 1: ***Average carbon intensity and average daily variability for 123 regions around the globe***



(a) Up to 10 weeks slack (b) Fixed Slack (24h)

Figure 2: ***Savings from Temporal Shifting.***

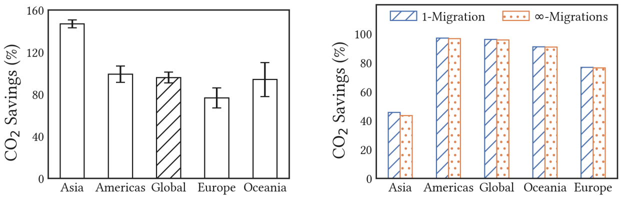
batch jobs (up to a week), based on Google’s Borg cluster traces [3]. The temporal flexibility of a job is determined by its *slack*, with deferrable jobs unable to be paused once started, while interruptible jobs can be paused and resumed.

3 CARBON’S MAGNITUDE AND VARIANCE

Since spatiotemporal workload shifting exploits differences in grid carbon intensity across time and regions, we analyze carbon traces from 123 regions to understand their average carbon intensity and daily variation–coefficient of variation (CV). Figure 1 depicts the average and CV of each region. A significant number (54%) of the regions have below average (< 400 g·CO$_2$e/kWh) carbon intensity. This implies that workloads from the top two quadrants will benefit from spatial workload shifting by migrating workloads to cloud regions in the bottom two quadrants. The figure also reveals that a majority (57%) of regions have a below-average CV. Thus, overall few regions will see significant benefits from temporal shifting methods. Globally, carbon intensity has a medium average magnitude but varies widely, with a relatively low average variance.

4 SPATIOTEMPORAL ANALYSIS

Temporal: Deferrability + Interruptibility Figure 2(a) shows the potential global carbon savings given that a job can be delayed up to 10 weeks. Globally, the savings range from 10-30%. Oceania has a higher renewable penetration and more variation, resulting in carbon savings ranging from 25-50%. Figure 2(b) shows the average carbon savings from deferrability and interruptibility for a fixed slack of 24hrs for all 123 regions. The savings are relative to the global average carbon intensity of 368.39 g·CO$_2$eq/kWh. It can be seen that as the job lengths gradually increase, the savings from deferrability decrease. When the jobs are interruptible, the effect of interruptibility augments the savings for long jobs. Nevertheless, carbon savings for large jobs are limited compared to small jobs.



(a) Migrate to lowest globally (b) One and ∞ within grouping

Figure 3: ***Savings from Spatial Shifting.***

Spatial: One-Time Migration to Lowest Figure 3(a) shows the global average carbon savings. In this case, we assume an always-running job that migrates to the region with the lowest average carbon intensity for the entire year. In our trace, Sweden has the lowest annual average carbon intensity of ~16g·CO$_2$eq/kWh. Therefore, in the extreme case where every job from every region migrates to Sweden, the global average carbon savings are up to 95.68%. Moreover, there is an assumption that the carbon intensity patterns of different regions vary daily and thus may cross each other, so a sophisticated migration may be needed to extract more carbon savings rather than a simple one-time migration policy. To evaluate this hypothesis, we devise another migration policy called ∞-migrations since it can migrate an infinite number of times, as compared to the 1-migration policy we have used so far. Figure 3(b) shows the amount of relative carbon savings for the ∞-migration and 1-migration policies. The results show that carbon savings for both policies are quite similar, with ∞-migration yielding only slightly higher savings but the difference between them being <1%. Thus, migrating once to the greenest region yields the vast majority of the savings, and more sophisticated migration approaches that migrate more often are not necessary. Notably, our ∞-migration represents a best-case policy, as it ignores the overhead of migration. Thus, any practical policy that outperforms 1-migration would have a very tight upper bound on its savings.

5 CONCLUSIONS

We conducted an empirical analysis of the benefits and limitations of spatiotemporal workload shifting in the cloud. Our results show that although there is the potential for some significant carbon savings from spatiotemporal workload shifting, the benefits are often limited in practice. For temporal shifting, these limits derive from a lack of variability in carbon intensity at many locations. The locations where reducing carbon emissions is most important tend to be those with the highest absolute carbon emissions. On the contrary, locations with significant variability already have low average carbon emissions, so temporal shifting yields limited savings. While a simple spatial shifting gives substantial carbon savings, resource constraints and migration overheads will likely prevent jobs from migrating to the lowest carbon regions, limiting the benefits of spatial shifting.

REFERENCES

[1] C. Garcia. 2022. AKCP, The Real Amount of Energy A Data Center Uses. https://www.akcp.com/blog/the-real-amount-of-energy-a-data-center-use/.

[2] Kasper Groes Albin Ludvigsen. 2022. The Carbon Footprint of ChatGPT. https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d.

[3] 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 *European Conference on Computer Systems (EuroSys)*. ACM, New York, NY, USA, 1–14.