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

Thanathorn Sukprasert
University of Massachusetts Amherst
USA

Abel Souza
University of Massachusetts Amherst
USA

Noman Bashir
Massachusetts Institute of Technology
USA

David Irwin
University of Massachusetts Amherst
USA

Prashant Shenoy
University of Massachusetts Amherst
USA

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."

CCS Concepts: • **General and reference → Performance**; • **Social and professional topics → Sustainability**; • **Computer systems organization → Cloud computing**.

Keywords: Sustainable computing, carbon-aware workload optimizations, carbon footprint, cloud computing

1 Introduction

The demand for computing continues to increase rapidly and is expected to accelerate further with the mainstream adoption of machine learning (ML) and artificial intelligence (AI) applications, such as ChatGPT [10] and its derivatives. Since computation requires energy, computing's energy consumption is also expected to accelerate in the coming decades. For example, recent estimates project that datacenter energy consumption will increase by at least 10% per year until 2030 [6], which is significantly higher than the 1.65% estimated increase per year in the 2010s [30]. Given these trends, there is an increasing concern that this substantial growth in computing's energy consumption will lead to a proportionate increase in its carbon emissions. Technology companies have recognized this problem and are addressing it by setting aggressive targets for reducing computing's carbon footprint, e.g., such as achieving net-zero emissions by 2030 or even earlier [2, 15, 34, 36, 42, 43].

To achieve the aggressive carbon reduction goals above, researchers have begun to focus on optimizing computing's *carbon-efficiency*, or computations per unit of carbon emitted [8, 44], in addition to its energy efficiency, or computations per joule of energy consumed. While optimizing for energy efficiency reduces carbon emissions, the benefits will likely be limited moving forward as computing is already highly energy-efficient. Thus, to optimize carbon-efficiency, recent work has focused on leveraging real-time variations in energy's carbon-intensity across time and space by shifting computation to when and where lower-carbon energy is available. The recent emergence of third-party carbon information services [29, 48], which provide real-time data on energy's carbon-intensity at high temporal and spatial resolution, have enabled this approach. Cloud platforms have used these services to develop tools that provide coarse, high-level per-region estimates of energy's carbon-intensity [24]. This has led recent work to propose a range of spatial and temporal workload shifting policies that leverage energy's carbon-intensity variations to reduce computing's carbon emissions [1, 19, 22, 35, 39, 44, 49, 51].

To illustrate, Figure 1(a) shows that grid energy's carbon-intensity can vary by 2× over a day (in California) and by over 43× across regions (between Ontario and Mumbai). The magnitude and variability of grid energy's carbon-intensity depend on the mix of energy sources. Traditional fossil fuel-based energy sources, such as coal and natural gas, tend to exhibit high carbon-intensity with low variance. In contrast, renewable sources like solar and wind have low carbon

*This work will appear at EuroSys'24.