Decarbonizing the Petaflop | Research Statement

Abel Souza
asouza.io

Cloud datacenters play a vital role as the de facto environment for executing and testing scientific models, simulations and online services that support research in climate change, chemistry, web services, among others. As society explores new applications involving computing, exemplified by the widespread adoption of Artificial Intelligence (AI) and connected devices, millions of services are now delivered by datacenters through the Internet of Things (IoT). Although this transformative era will witness datacenters handling yottabyte-scale data ($\sim$1 trillion terabytes), this increase in demand is raising serious concerns about its long-term environmental sustainability. Today, datacenters and networks consume approximately 3% of the world's electricity [1], with projections indicating a five-fold increase by 2030. However, contrary to intuition, the core issue is not the escalation of energy consumption itself, but the carbon footprint resulting from such energy usage. In the past decades, and even today, most efforts to promote sustainable computing primarily revolve around enhancing energy efficiency, with limited results in terms of actual reductions in carbon. Nevertheless, with demand surpassing gains in efficiency, unless a different approach is devised, computing is poised to become one of the largest contributors to greenhouse gas emissions (GHG) [2, 3].

I specialize in researching large-scale distributed systems, focusing on the optimization of resource management. Recently, I have targeted the sustainability aspects within these systems by exploring strategies that elevate *carbon efficiency* as a design metric. My main argument centers on the idea that platforms should expose the energy system, granting applications the visibility and control necessary to create their own energy and resource management policies based on their unique requirements. In pursuit of these objectives, I have made fundamental contributions in control and optimization, and further integrated these in real system implementations:

- I've explored practical abstractions to manage constraints for achieving sustainable operations, seamlessly integrating these innovations into real-world system implementations. Moreover, I've addressed limitations associated with these abstractions and proposed various techniques to enhance their effectiveness (§1).
- In my research, I have created novel resource management mechanisms and policies to enhance system design. As an experimental system's researcher, I engage in the development, optimization, and evaluation of software artifacts to gain insights into the performance of large-scale systems (§2).
- Simultaneously, I'll continue to employ rigorous theoretical approaches, including analytical models, optimization, and control techniques to comprehensively understand and optimize the systems I construct (§3).

1 Carbon-Aware Computing for Greener Operations

Intuition The ongoing shift of information and communication technology operations to hyperscale datacenters brings several advantages, including on-demand scaling and pay-as-you-go pricing. As a consequence, this wide scale increase in usage is escalating the energy needs of cloud datacenters. Thus, it is imperative to accurately account for and address the factors contributing to the associated environmental consequences. The way forward lies in enabling all levels of the software stack – operators, applications, appliances, and users – to adapt to the availability of unreliable low-carbon energy, a capability currently unexplored due to the lack of visibility into the information associated with energy consumption [2, 3, 4, 5, 6].

Energy Hypervisor

Over the last thirty years, considerable efforts have been made to enhance energy efficiency and management in datacenters [2]. Unlike energy-efficient systems that involves optimizing the energy use of internal components, designing *carbon-efficient* systems involves examining the local energy system and grid to understand the sources and characteristics of energy. In addition, research has mainly focused on virtualizing and optimizing computational resources, with no work on enabling control of the physical energy system supporting applications [2]. To address these challenges, I introduced the Ecovisor [4], an innovative framework that virtualizes the energy system, rather than resources. This allows each application to have **software-defined control over energy**, enabling them to handle the unreliability of clean energy based on their individual requirements. We have demonstrated the flexibility of Ecovisor by implementing a variety of applications that utilize its API in different ways to optimize carbon-efficiency. For instance, we have implemented an elastic ML training system that adjusts resource usage to cap carbon emissions, resulting in a reduction of up to 22% in footprint without sacrificing performance. Furthermore, this research is one of the main efforts in the NSF funded project CarbonFirst, and has resulted in securing additional funds of over $15,000, supporting the efforts of our research laboratory in sustainable computing.

Carbon-Aware Scheduling

Modern distributed applications are naturally well-positioned to reduce their carbon emissions due to their fault-tolerance and load balancing characteristics that inherently incorporate numerous features that can adeptly handle the key intermittency issue associated with renewable energy sources. Moreover, computation has significant spatial and temporal flexibility, enabling it to shift time, location, and intensity of execution to better align it with the availability of carbon-free renewable energy or low-carbon energy [2, 3]. In evaluating temporal versus spatial flexibility, it is crucial to calculate the actual benefits, limitations, and overheads for real cloud workloads. My research aims at directly confronting these issues by evaluating the use of carbon as a fundamental metric in computing platforms to effectively signal the impacts of GHG emissions. In a recent work [5], I analyzed the energy profile of hundreds of countries to evaluate the resulting carbon reductions when shifting computations over space and time. The analysis revealed that although workloads can reduce their carbon emissions via spatiotemporal workload shifting, the practical upper bounds are limited and not ideal, highlighting the need of new adaptive techniques and systems to cope with such limitations. As one of such techniques, my recent research has designed CASPER, a carbon-aware scheduler for distributed web applications [6]. My technique underscored the substantial potential to reduce the carbon footprint of geo-distributed cloud applications, with carbon savings of up to 70% and no performance losses. This approach stands as a key development in carbon-aware schedulers for distributed applications due to its feasibility in carbon savings.

2 Designing Elastic Cloud Datacenters

Intuition Modern cloud workloads exhibit dynamic and evolving characteristics. This involves fluctuations in demand, varying resource requirements, and the ability to scale both horizontally and vertically based on changing needs. Orchestrators and operating systems abstract the resource allocation from applications to simplify management, but this limits applications ability to adapt to variations in user requirements. As such, to optimize cloud-efficiency, resource managers and operating systems require an improved ability to measure and control applications own performance. My past research has explored system abstractions and interfaces to enable new levels of efficiency and novel use cases within highly constrained and complex infrastructures, including in cloud [7, 8] and HPC [9, 10, 11].

Improving System Level Scheduling

In distributed and multi-tiered applications, various inefficiencies, such as operating system delays, request reordering, and network congestion, contribute to extended runtimes and latency. While recent research has aimed

at reducing delays for individual application components, like through replication and scheduling, my past work advocates for comprehensive approaches that effectively minimize end-to-end makespans and tail latency across all components of an application. For batch workloads, our research demonstrates enhancements in resource utilization and performance, resulting in increased overall efficiency. This improvement allows for accomplishing more with the same capacity [9]. For interactive workloads I introduced TailTamer [8], a novel OS scheduler that strategically organizes user requests based on their original arrival time, countering performance degradation caused by request reordering common in multi-tiered distributed applications, ultimately reducing the response time. TailTamer operates as a userspace Linux scheduler, exerting control over the operating system. In comparison against state-of-the-art schedulers such as Linux and Shinjuku [12], TailTamer significantly improves the response time by up to $3\times$, while greatly enhancing the overall amount of tasks the system can perform under resource constraints. This multi-stack control enables the systems I construct to operate as close to their theoretical limits as possible.

Autonomic Management

As technology trends evolved towards data-intensive ML workloads, researchers started reevaluating system designs built on old decades assumptions of non-adaptive workloads. The goal of my early research was to minimize the negative impact on application requirements and Quality-of-Service (QoS) while improving overall datacenter efficiency and utilization. It focused on proposing and evaluating autonomic solutions [9], novel scheduling strategies [8, 10, 11], and software architectures to enhance the performance and efficiency of new data-intensive workloads with zero impacts on traditional loads [9, 13]. These approaches were designed to seamlessly support a diverse range of applications across different real-world infrastructure and operational scenarios, including anomalous behaviors such as hardware failures and transiency and tightened deadlines.

Augmenting User Applications

Aligned with these objectives, my early research also aimed to enable users to specify application requirements, thereby enhancing the system utilization and applications efficiency, such as scientific workflows commonly found in HPC [13]. We have developed theoretical control techniques for adapting and selecting optimal fault-tolerant mechanisms based on workload variations and application characteristics [7, 14], and evaluated machine learning and reinforcement learning methods for autonomic capacity control, continuously improving application performance [11]. These developed methods not only simplified and improved system operations, but also enable the introduction of novel workflows and applications that were previously unsupported.

3 Future work

Intuition The limited scope for further optimizations in energy-efficiency [3] demand that the cloud, edge, and cyber-physical systems are considered as a spectrum encompassing computational and spatiotemporal characteristics. Based on works such as Ecovisor and CASPER, my future endeavors will investigate the development of resilient distributed system that support the digital continuum encompassing data intensive workflows in IoT, cloud-edge computing and cyber-physical systems, treating energy and carbon footprint as fundamental objectives. While addressing the constraints imposed by platform-level limits, my future research aims to achieve a harmonious balance between empowering users and optimizing system performance [15]. This will be accomplished through innovative platform designs and mechanisms that overcome existing limitations.

Greening Edge and Cloud Datacenters

Despite the imperative for all of cloud energy consumption to originate from low-carbon sources, the inherent difficulty arises when user applications inadvertently take coordinated actions in bulk, leading to a negative impact on both the cloud and grid systems, commonly referred as the *stampede effect* or spikes. In scenarios where

multiple workloads exhibit correlated resource usage peaks, the cloud may also face capacity shortages during peak times, while experiencing significant underutilization during off-peak periods. The cumulative effect of each job independently deciding to optimize its own carbon emissions inadvertently leads to increased overall carbon emissions because the cloud will provision additional nodes to accommodate fluctuations in demand. In the future, I plan to explore higher-level layers to present operators with a comprehensive view of emissions across the cloud. The approach can employ a black-box approach transparent to the applications and use anticipated utilization profile, returning a score to indicate the degree of alignment between the profile and the platform's carbon preferences. The score encompasses various metrics, including carbon optimization with workload profiling and forecasting, as effectively applied in previous work [6, 7]k requiring unique approaches to optimize for carbon efficiency. To this end, we will develop higher-level system abstractions that facilitate carbon-efficient applications at each scale.

Exploiting Performance Flexibility

Present-day resource management solutions advocate a model where users specify a particular quantity of resources, a range, or even refrain from specifying them at all. While this model declares the end cluster state through an imperative definition of requirements and has demonstrated effectiveness, it fails to incorporate the dynamics of applications from users objectives, limiting potential optimization opportunities. For instance, one issue relates to the suboptimal requirement specification that results in resource overallocation and underutilization [9]. Moreover, these declarations carry a contextual requirement that users often struggle to comprehend [13, 10]. In the future, I will propose a paradigm shift in orchestration: ***transitioning from an imperative model to an intent-driven model***. In this paradigm, users specify intentions through objectives – e.g., a SLA targets such as latency, throughput, power or carbon –, and the orchestration stack itself determines the necessary resources to achieve it. This approach will build upon ongoing community efforts in scheduling that determines when and where to place workloads, and be complemented by a continuous planning layer that decides what and how to configure intents in the system. The primary objective is to concentrate on overseeing the Quality-of-Service (QoS) of a specific set of workload instances. Use cases include Augmented Reality (AR), signal and image processing, machine learning, and HPC simulations. Through this research, we aim to contribute to the advancement of more efficient and sustainable computing by providing innovative solutions for managing energy and carbon in cloud-edge systems.

Human-Centered Carbon Management

The success of Electric Vehicles (EVs) surpassing that of internal-combustion engine vehicles becomes less relevant if, ultimately, these EVs are charged with energy derived from fossil-fuel-based power plants. Similarly, this reasoning can be extended to apply to all electrification and digitalization projects that have been recently announced, whereas major industries have committed to reducing their carbon emissions, primarily by improving energy efficiency and by offsetting their footprint through power purchase agreements. However, there is a lack of standard controllers and software interfaces empowering users to manage their energy consumption based on their specific needs. In my future research, I will work on the development of distributed models that enable users the full control of household appliances, extending the principles proposed in the Ecovisor to various aspects of daily life that possess highly temporal flexibility – e.g., laundry, dish-washing and heating –, all while coordinating with the grid to avoid saturating it. Additionally, my plan involves investigating the impact and quantification of sustainability initiatives through a direct human-in-the-loop approach, incorporating user design, participation, and feedback. For instance, it is possible to develop behavioral usage models based on sensor activation in homes to capture *actual* usage patterns to properly schedule household appliances use [15]. To initiate this process, we have obtained approval from an institutional review board (IRB) that will be utilized to securely gather extensive data for impact analysis, user-modeling and validation. Beyond this, additional challenges reside in incorporating such a solution in city-scale environments.

References

[1] Vida Rozite, Emi Bertoli, and Brendan Reidenbach. Data centres and data transmission networks. *IEA*, 2023.

[2] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, Irwin David, Prashant Shenoy, Ramesh Sitaraman, **Abel Souza**, and Adam Wierman. Enabling sustainable clouds: The case for virtualizing the energy system. In *Proceedings of the ACM Symposium on Cloud Computing*, 2021.

[3] Noman Bashir, David Irwin, Prashant Shenoy, and **Abel Souza**. Sustainable computing-without the hot air. *ACM SIGENERGY Energy Informatics Review*, 2023.

[4] **Abel Souza**, Noman Bashir, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. Ecovisor: A virtual energy system for carbon-efficient applications. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

[5] Thanathorn Sukprasert, **Abel Souza**, Noman Bashir, David Irwin, and Prashant Shenoy. Quantifying the benefits of carbon-aware temporal and spatial workload shifting in the cloud. *arXiv preprint arXiv:2306.06502*, 2023.

[6] **Abel Souza**, Jasoria Shruti, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. Casper: Carbon-aware scheduling and provisioning for distributed web services. In *Proceedings of the 14th International Green and Sustainable Computing Conference (IGSC), Toronto, ON, Canada*, 2023.

[7] **Abel Souza**, Alessandro Vittorio Papadopoulos, Luis Tomas, David Gilbert, and Johan Tordsson. Hybrid adaptive checkpointing for virtual machine fault tolerance. In *2018 IEEE International Conference on Cloud Engineering (IC2E)*, 2018.

[8] Nathan Ng, **Abel Souza**, Ahmed Ali-Eldin, Cristian Klein, David Irwin, Don Towsley, and Prashant Shenoy. Tailtamer: Reducing tail response time through system-wide scheduling. *Submitted*, 2024.

[9] **Abel Souza**, Mohamad Rezaei, Erwin Laure, and Johan Tordsson. Hybrid Resource Management for HPC and Data Intensive Workloads. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, 2019.

[10] **Abel Souza**, Kristiaan Pelckmans, Devarshi Ghoshal, Lavanya Ramakrishnan, and Johan Tordsson. ASA – The Adaptive Scheduling Architecture. In *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, 2020.

[11] **Abel Souza**, Kristiaan Pelckmans, and Johan Tordsson. A HPC Co-Scheduler with Reinforcement Learning. In *24th International Workshop on Job Scheduling Strategies for Parallel Processing*, 2021.

[12] Kostis Kaffes, Timothy Chong, Jack Tigar Humphries, Adam Belay, David Mazières, and Christos Kozyrakis. Shinjuku: Preemptive scheduling for $\{\mu$second-scale$\}$ tail latency. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, 2019.

[13] William Fox, Devarshi Ghoshal, **Abel Souza**, Gonzalo P Rodrigo, and Lavanya Ramakrishnan. E-HPC: A Library for Elastic Resource Management in HPC Environments. In *Proceedings of the 12th Workshop on Workflows in Support of Large-Scale Science*, 2017.

[14] Felipe Gutierrez, Kaustubh Beedkar, **Abel Souza**, and Volker Markl. Adcom: Adaptive combiner for streaming aggregations. In *EDBT 2021-24th International Conference on Extending Database Technology*, 2021.

[15] **Abel Souza**, Mihir Shenoy, and Camellia Zakaria. Empowering user-centered carbon management: Bridging individual preferences and sociotechnical advancements. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2023.