Environmentally-Conscious Cloud Orchestration Considering Geo-Distributed Data Centers

Giulio Attenni, Novella Bartolini
{attenni,bartolini}@di.uniroma1.it
Sapienza University of Rome, Italy

Abstract

This paper presents a framework for environmentally-conscious job deployment and migration in cloud environments, aiming to minimize the environmental impact of resource provisioning while incorporating sustainability requirements. As the demand for sustainable cloud services grows, it is crucial for cloud customers to select data center operators based on sustainability metrics and being able to report the ecological footprint of their services. We formalize the problem and propose an efficient algorithm for its optimal solution. Additionally, we outline potential future works where more constrained problems can be considered.

Keywords

Data center sustainability, Cloud orchestration, Carbon-aware

1 Introduction

The demand of cloud services is set to rise in the coming years, and the liked environmental impact represents a critical concern. Data center sustainability is gaining more and more attention, and the need of comprehensive measures to reduce the environmental impact is becoming clear. In response to the growth of data center economy, the European regulatory landscape rapidly evolved. The JRC [9] in 2018 established the voluntary European Code of Conduct for Data Centers [8]. Then, in 2021, the Climate Neutral Data Center Pact was launched, to make data centers and cloud infrastructure services in Europe climate-neutral by 2030. Signatories pledge to meet quantifiable goals, such as using 100% renewable energy while also making recycling and water conservation a priority [1]. Moreover, from 2024 the Energy Efficiency Directive's sets reporting obligations for data centers with a power demand of at least 500kW [7]. To enhance and standardize sustainability reporting among companies in 2023 the Corporate Sustainability Reporting Directive has been enacted, which imposes that companies must disclose detailed information regarding their environmental and social impacts, sustainability risks, and governance practices [6]. Finally, in march 2024 the EU Commission has adopted a new delegated regulation for establishing an EU-wide scheme to rate the sustainability of data centers [5]. Thus, both for data center operators and companies relying on cloud intensive computations that need to meet their environmental pledges, it is crucial to consider the environmental footprint of their operations in corporate sustainability reports. We propose an environmentally-conscious cloud orchestrator that optimizes job provisioning and migration to minimize the environmental footprint, taking into account users' requirements and the need to comply with sustainability standards for data centers.

2 Related Work

Enhancing resource management and operational efficiency, cloud orchestration has fueled a wide body of research [22]. Allocating resources, both optimizing energy consumption [10; 18; 20; 23] and carbon awareness [2; 4; 10; 12; 17; 19; 23] have received a significant interest. Stojkovic et al. [20] propose EcoFaaS, an energy management framework designed for serverless environments which optimizes the overall energy consumption. Rastegar et al. [18] propose an LP relaxation for energy-aware execution scheduler for serverless service providers, minimization of energy consumption for executing the incoming chains of functions with specified computational loads and deadlines. Gao et al. [10] provide a scheduler that controls the traffic directed to each data center optimizing a 3way-trade-off between access latency, electricity cost, and carbon footprint. Zhou et al. [23] use Lyapunov optimization to perform load balancing across geo-distributed data centers; capacity rightsizing; and server speed scaling. The objective here is to optimize 3way-trade-off between electricity cost, SLA (Service-Level Agreement) requirement and emission reduction budget. Souza et al. [19] propose a provisioner that minimizes both the number of active servers and the associated carbon emission. Maji et al. [12] propose load balancing in VMware's Avi Global Server Load Balancer, which uses a linear scoring function to select the optimal data center in terms of marginal carbon intensity and the distance between the client and the data center. Cordingly et al. [4] propose a prototype for computing resource aggregation that minimizes the carbon footprint of a serverless application through carbon-aware load distribution. Piontek et al. [17] propose a Kubernetes scheduler which shifts non-critical jobs in time so to reduce carbon emissions based on their prediction algorithm. Chadha et al. [2] present GreenCourier, a Kubernetes scheduler designed to reduce carbon emissions associated to serverless functions scheduled across geographically distributed regions.

Differently from previous works, we aim to consider sustainability as a whole, accounting for more than energy consumption and the related GHG (greenhouse gas) emissions. We aim to propose a holistic approach that encompasses the impact of the whole production cycle.

3 Problem Formulation

In this section, we formally define the problem of job deployment and migration considering sustainability profiles of data center geographically located in several regions and user sustainability

LOCO 2024, 3 December 3034, University of Glasgow

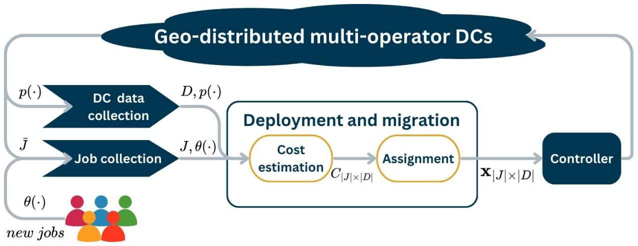

Figure 1: Architecture

preferences, that may derive from the need to adhere to sustainability standards and laws. Data center profiles represents various environmental impact factors and user requirements prioritize over the impact factors.

3.1 Environmental impact factors

Data centers' environmental impact can be categorized into two broad areas: operational and construction/dismantling impacts. Each phase of a data center's life cycle contributes to its ecological footprint.

The ongoing operation of data centers is resource-intensive, particularly in terms of energy consumption and water usage. Several metrics can be considered to account for operational impact [21]. The impact of energy production has to be considered, since data centers are among the most energy-demanding facilities globally. They require large amounts of electricity to power servers, cooling systems, and backup power supplies. It is estimated that data centers in 2022 accounted for nearly 500 TWh, and it is expected to rise above 800 TWh of consumption in 2026, with cloud-based services, 5G networks, AI and cryptocurrencies being the major drivers of such demand increase [24]. This energy usage results in significant GHG emissions, especially in regions where electricity is generated from fossil fuels. However, GHG emissions is not the only factor that varies with the energy source. Namely: *water footprint*, which contributes to environmental pollution; *land use*[14], which affects ecosystems and biodiversity; *deathprint* [3; 15], which is the number of people killed per kWh produced; *waste*, which contributes to environmental pollution; *critical raw metals* and *material use*, which affect resource depletion. The environment impact of data centers is directly tied to the energy mix of their respective grids. Moreover, cooling systems have a huge impact on the environment, since they are both energy-intensive and water-intensive. Indeed, data centers rely on vast amounts of water for cooling purposes. It is estimated that a typical mid-sized data center can use approximately 25.5 million liters of water each year when employing traditional cooling methods [11]. This heavy water reliance can exacerbate local water scarcity, especially in regions prone to drought. Example of metrics related to cooling systems are: the Air Economizer Utilization Factor and the Water Usage Effectiveness [21].

Data center facilities have a significant environmental impact also for what concerns the construction and eventual dismantling. Large amounts of raw materials like concrete, steel, copper, and rare earth elements are required for its construction. Moreover, these facilities also occupy significant land. The level of the ecological footprint of a data center building can be formally certified, e.g. with the LEED certification scheme [13]. The dismantling of data centers generates substantial e-waste, which often contains toxic materials if not properly recycled, can leach into soil and water, and ultimately can have an adverse impact on human health and the environment [16].

3.2 Deployment and Migration

Let us consider $\mathcal{F} = \{1, ..., F\}$ the set of environmental impact factors, $J = \{1, ..., M\}$ the set of jobs, $U = \{1, ..., K\}$ the set of users, and $D = \{1, ..., N\}$ the set of Data Centers. Moreover, let us consider the following constant functions: $u : J \longrightarrow U$, which is the function providing the owners of each job; $\theta : U \longrightarrow \mathbb{R}^F$, which is the function that provides the scores to prioritize factors of interest to meet the user requirements; $p : D \longrightarrow \mathbb{R}^F$, which is the function that provides the data centers environmental profiles, i.e. scores which rate their impact on each environmental impact factor; $d : \bar{J} \longrightarrow D$ where $\bar{J} \subseteq J$ is the subset of already deployed jobs and the function d provides the data center on which each job is deployed; and $\mathbb{1} : J \longrightarrow \{0, 1\}$ returns 1 if its argument belongs to $\bar{J}$, 0 otherwise. Moreover, let us consider the cost of deploying a job in a certain data center $c_d : D \longrightarrow \mathbb{R}$, and the migration cost of a job $c_m : D \times D \longrightarrow \mathbb{R}$ (note that for each $n \in D$, $c_m(n, n) = 0$). Finally, let us define a binary decision variable for each $m \in J$ and $n \in D$ which indicates whether the m-th job is deployed on the n-th data center. Namely, $x_{m,n} = 1$ if the m-th job is deployed on the n-th data center and $x_{m,n} = 0$ otherwise. Now, let us define the cost of the decision $x_{m,n}$ as follow: $C(m, n) = p(n)^T \theta(u(m)) + \mathbb{1}(m) \cdot c_m(d(m), n) + (1 - \mathbb{1}(m)) \cdot c_d(n)$.

The first factor accounts for the execution cost considering the user requirements and the data center profile, the second factor accounts for the migration costs whether a job is migrated, and finally the third factor accounts for the deployment cost whether it is the first time the job has to be deployed. Finally, we can define the following optimization problem:

$$\min \sum_{m\in J} \sum_{n\in D} C(m, n) \cdot x_{m,n} \tag{1}$$

$$\text{s.t.} \sum_{n\in D} x_{m,n} = 1 \qquad \forall m \in J \tag{2}$$

$$x_{m,n} \in \{0, 1\} \qquad \forall n \in D, m \in J \tag{3}$$

To solve this problem, we propose an algorithm that finds the optimal solution in $O(MN)$. Namely, for each job m we assign the data center n for which the assignment cost $C(m, n)$ is minimum. This problem becomes NP-hard as we consider more constraints such as data center service capacity or user cost budgeting. Thus, heuristic approaches might be necessary to handle such problem.

4 Conclusion and future works

In this extended abstract, we tackle the challenge of job deployment and migration to optimize resource provisioning while minimizing environmental impact, incorporating customer preferences and a holistic approach to sustainability. This framework will help cloud customers select data centers based on environmental impact and report their ecological footprint. Future work includes formalizing data center sustainability profiles, focusing on open data, conducting extensive simulations, and exploring a more constrained version of the problem.

References

[1] Climate neutral data centre pact, 2024. Accessed: 2024-09-16. URL: https://www.climateneutraldatacentre.net.

[2] Mohak Chadha, Thandayuthapani Subramanian, Eishi Arima, Michael Gerndt, Martin Schulz, and Osama Abboud. Greencourier: Carbon-aware scheduling for serverless functions. In *Proceedings of the 9th International Workshop on Serverless Computing*, WoSC '23, page 18–23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3631295.3631396.

[3] James Conca. Energy's deathprint: A price always paid, 2012. Accessed: 2024-09-16. URL: https://www.forbes.com/sites/jamesconca/2012/06/10/energys-deathprint-a-price-always-paid/.

[4] Robert Cordingly, Jasleen Kaur, Divyansh Dwivedi, and Wes Lloyd. Towards serverless sky computing: An investigation on global workload distribution to mitigate carbon intensity, network latency, and cost. In *2023 IEEE International Conference on Cloud Engineering (IC2E)*, pages 59–69, 2023. doi:10.1109/IC2E59103.2023.00015.

[5] European Commission. Commission adopts eu-wide scheme for rating sustainability of data centres, 2024. Accessed: 2024-09-16. URL: https://energy.ec.europa.eu/news/commission-adopts-eu-wide-scheme-rating-sustainability-data-centres-2024-03-15_en.

[6] European Commission. Corporate sustainability reporting, 2024. Accessed: 2024-09-16. URL: https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en.

[7] European Commission. Energy efficiency directive, 2024. Accessed: 2024-09-16. URL: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en.

[8] European Commission Joint Research Centre. European code of conduct for energy efficiency in data centres, 2022. Accessed: 2024-09-16. URL: https://joint-research-centre.ec.europa.eu/scientific-activities-z/energy-efficiency/energy-efficiency-products/code-conduct-ict/european-code-conduct-energy-efficiency-data-centres_en.

[9] European Commission Joint Research Centre. Joint research centre, 2024. Accessed: 2024-09-16. URL: https://joint-research-centre.ec.europa.eu/index_en.

[10] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. It's not easy being green. In *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '12, page 211–222, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2342356.2342398.

[11] H2O Building Services. How much water do data centres use?, 2024. Accessed: 2024-09-16. URL: https://smartwatermagazine.com/news/h2o-building-services/how-much-water-do-data-centres-use.

[12] Diptyaroop Maji, Ben Pfaff, Vipin P R, Rajagopal Sreenivasan, Victor Firoiu, Sreeram Iyer, Colleen Josephson, Zhelong Pan, and Ramesh K Sitaraman. Bringing carbon awareness to multi-cloud application delivery. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, HotCarbon '23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3604930.3605711.

[13] NetZero Events. What are the sustainable data center standards and certification?, 2024. Accessed: 2024-09-16. URL: https://netzero-events.com/what-are-the-sustainable-data-center-standards-and-certification/.

[14] Our World in Data. Land use per energy source, 2024. Accessed: 2024-09-16. URL: https://ourworldindata.org/land-use-per-energy-source.

[15] Our World in Data. The safest sources of energy, 2024. Accessed: 2024-09-16. URL: https://ourworldindata.org/safest-sources-of-energy.

[16] Violet N Pinto. E-waste hazard: The impending challenge. *Indian journal of occupational and environmental medicine*, 12(2):65–70, 2008.

[17] Tobias Piontek, Kawsar Haghshenas, and Marco Aiello. Carbon emission-aware job scheduling for kubernetes deployments. *The Journal of Supercomputing*, 80(1):549–569, 2024.

[18] Seyed Hamed Rastegar, Hossein Shafiei, and Ahmad Khonsari. Enex: An energy-aware execution scheduler for serverless computing. *IEEE Transactions on Industrial Informatics*, 20(2):2342–2353, 2024. doi:10.1109/TII.2023.3290985.

[19] Abel Souza, Shruti Jasoria, 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 '23, page 67–73, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3634769.3634812.

[20] Jovan Stojkovic, Nikoleta Iliakopoulou, Tianyin Xu, Hubertus Franke, and Josep Torrellas. Ecofaas: Rethinking the design of serverless environments for energy efficiency. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 471–486, 2024. doi:10.1109/ISCA59077.2024.00042.

[21] Sunbird DCIM. Top 30 data center sustainability metrics, 2024. Accessed: 2024-09-16. URL: https://www.sunbirddcim.com/infographic/top-30-data-center-sustainability-metrics.

[22] Denis Weerasiri, Moshe Chai Barukh, Boualem Benatallah, Quan Z. Sheng, and Rajiv Ranjan. A taxonomy and survey of cloud resource orchestration techniques. *ACM Comput. Surv.*, 50(2), may 2017. doi:10.1145/3054177.

[23] Zhi Zhou, Fangming Liu, Ruolan Zou, Jiangchuan Liu, Hong Xu, and Hai Jin. Carbon-aware online control of geo-distributed cloud services. *IEEE Transactions on Parallel and Distributed Systems*, 27(9):2506–2519, 2016. doi:10.1109/TPDS.2015.2504978.

[24] Çam et al. Electricity 2024, 2024. Accessed on September 9, 2024. URL: https://www.iea.org/reports/electricity-2024.