Green Cloud Computing: Challenges and Opportunities

Daniel Cordeiro[1], Emilio Francesquini[2], Marcos Amarís[3], Márcio Castro[4], Alexandro Baldassin[5], João V. F. Lima[6]

[1] Universidade de São Paulo (USP), São Paulo, Brasil

[2]Universidade Federal do ABC (UFABC), Santo André, Brasil

[3]Universidade Federal do Pará (UFPA), Tucuruí, Brasil

[4]Universidade Federal de Santa Catarina (UFSC), Florianópolis, Brasil

[5]Universidade Estadual Paulista (UNESP), Rio Claro, Brasil

[6]Universidade Federal de Santa Maria (UFSM), Santa Maria, Brasil

daniel.cordeiro@usp.br, e.francesquini@ufabc.edu.br, amaris@ufpa.br, marcio.castro@ufsc.br, alex@rc.unesp.br, jvlima@inf.ufsm.br

Abstract. *We discuss the immediate need, challenges, and opportunities to transition into greener cloud computing platforms. Actions must be taken not only by the academy and the industry, but also by computer science practitioners from several areas, such as information systems, scheduling theory, distributed systems, HPC, computer architecture, and approximate computing, to cite a few.*

1. Introduction

Cloud Computing (CC) has changed the industry, minimizing IT infrastructure costs, easing the deployment of new products, reducing maintenance costs, and rapidly adjusting resources to meet unpredictable demands [Oda et al. 2018]. Netflix, LinkedIn, Facebook, and other relevant industry players rely on CC to offer efficient and scalable solutions. Large-scale Data Centers (DC), distributed across different locations, are the foundation of CC. As the use of cloud-based solutions increases in popularity, so does the importance of energy efficiency and the reduction of the carbon footprint of large-scale DCs. Indeed, efforts to improve energy efficiency have been the focus of both academia and cloud providers. These improvements were the main factors responsible for limiting the energy consumption increase to only 6%, despite an estimated increase of 10× in IP traffic, 25× in storage capacity, and 6× in the workload capacity of DCs between 2010 and 2018 [Masanet et al. 2020]. Nevertheless, [Koot and Wijnhoven 2021] expect a combined growth of data center electricity needs of 286 TWh in 2016 up to 321 TWh in 2030, if today's technological and behavioral trends remain the same.

Maintaining an ever-increasing energy efficiency might not be feasible. Greenpeace reports [Greenpeace 2017] that, even if many DCs were already fully supplied with renewable energy in 2017, they are not the majority. In particular, DCs in Virginia (known to host 70% of US internet traffic) are nicknamed "Ground zero for the Dirty Internet" since they are powered by a mix of 2% of renewable energy and 37% coal. In this context, we propose a reflection on how we can make cloud applications on large-scale cloud infrastructures more sustainable by efficiently designing, managing, and optimizing their execution. In the next section, we focus on some challenges that researchers will need to address in the near future.

This work has been partially supported by grants #2019/26702-8 and #2021/06867-2, São Paulo Research Foundation (FAPESP), and by Federal University of Santa Catarina.

2. Challenges and Opportunities

Energy-aware Resource Management The dynamic nature of how CC platforms manage their resources enables different strategies to reduce energy consumption. Previous works showed that less energy can be used when Dynamic Speed Scaling is used in some of the machines [Cohen et al. 2014, Lima et al. 2019] or if migrations of virtual machines are employed to execute tasks on DCs with more renewable energy available [Vasconcelos et al. 2022]. Future energy-efficient, low-environmental impact DCs will need to decrease the usage of non-renewable energy. Particular focus will be needed on providing application developers the tools they need to maintain the scope of their development within their research subject and to help them to minimize the effort required to execute their applications in the cloud in a performance- and energy-efficient way. Cloud platforms are investing in the use of renewable energy. They are a suitable target for energy-efficient applications. However, the programmer is still responsible for efficiently utilizing the resources available (including specialized hardware) and simultaneously keeping the load balanced and the data coherent between the nodes while minimizing data movements among them.

Specialized Hardware and Software Crypto-mining is maybe one of the most well-known examples of specialized hardware being used to provide more efficient, performance and energy-wise solutions to a particular problem. In particular, it transitioned from regular CPUs to GPUs, FPGAs, and ASICs. However, additional improvements can also be achieved by specialized software that considers an application's intrinsic characteristics to provide faster solutions. Approximate Computing (AC) is an approach in which energy efficiency and better performance can be achieved by allowing certain approximations. For instance, when rendering a 3D model, the user might accept a slight inaccuracy in the color of a few pixels in exchange for speed. Recent work from our group [Rocha et al. 2022] showed that, by identifying and exploring recurring patterns during a large-scale metropolitan traffic simulation, it is possible to reduce the average processing time required to $\sim$50% (compared to a full simulation). This can be done while maintaining a high degree of accuracy (*e.g.*, with losses below 1.2% and 30% for average speed and average street occupancy rate estimates, respectively). In general, AC can be used whenever the precision provided by the computational system is superior to the level that the applications or their users actually need. Future DCs might become more efficient by providing specialized hardware or Software-as-a-Service (SaaS) solutions that consider each application's characteristics instead of a "one size fits all" solution.

High Performance Computing Cloud computing allows convenient on-demand access to a configurable group of computing resources rapidly with minimum effort and contact with the provider. In the High-Performance Computing (HPC) context, the benefits of using public cloud resources make it an attractive alternative to expensive on-premise HPC clusters. However, the software ecosystem necessary to make possible a sustainable HPC cloud platform is not yet mature. Cost advisors, large contract handlers, DevOps solutions, Application Programming Interfaces (APIs), and HPC-aware resource managers are current software gaps in this regard. We have been working on new solutions to efficiently manage and execute HPC scientific applications and workflows on public clouds, with sustainability as a common objective [Munhoz et al. 2022, Oda et al. 2018]. The proposed open-source tools and optimizations focus on multiple performance objectives: makespan (how to minimize applications' execution times), budget (how to choose virtual machine instances to meet the users' and applications' needs with cost savings), energy

(how to minimize energy consumption) as well as on fault tolerance (how to provide efficient fault tolerance to HPC applications running on cloud spot instances). Analytical models and machine learning algorithms have recently proved useful in predicting the execution time of HPC workloads [Amaris et al. 2023]. These solutions could be adapted to predict HPC applications' energy consumption and carbon footprint, helping researchers build new energy-aware schedulers.

3. Conclusion

We propose a discussion on how a transition into greener cloud computing platforms can be achieved. Some current, future challenges and efforts needed to make it a reality were outlined. However, it is already clear that much work remains to be done. This work will involve not only the academy and the industry but also computer science practitioners from several areas, such as information systems, scheduling theory, distributed systems, HPC, computer architecture, and approximate computing, to cite a few.

References

[Amaris et al. 2023] Amaris, M., Camargo, R., Cordeiro, D., Goldman, A., and Trystram, D. (2023). Evaluating execution time predictions on gpu kernels using an analytical model and machine learning techniques. *JPDC*, 171:66–78.

[Cohen et al. 2014] Cohen, J., Cordeiro, D., and Raphael, P. L. F. (2014). Energy-aware multi-organization scheduling problem. In *Euro-Par*, pages 186–197. Springer.

[Greenpeace 2017] Greenpeace (2017). Clicking Green: who is winning the race to build a green Internet. Greenpeace report.

[Koot and Wijnhoven 2021] Koot, M. and Wijnhoven, F. (2021). Usage impact on data center electricity needs: A system dynamic forecasting model. *Applied Energy*, 291:116798.

[Lima et al. 2019] Lima, J. V. F., Raïs, I., Lefèvre, L., and Gautier, T. (2019). Performance and energy analysis of openmp runtime systems with dense linear algebra algorithms. *IJHPCA*, 33(3):431–443.

[Masanet et al. 2020] Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986.

[Munhoz et al. 2022] Munhoz, V., Castro, M., and Mendizabal, O. (2022). Strategies for Fault-Tolerant Tightly-coupled HPC Workloads Running on Low-Budget Spot Cloud Infrastructures. In *IEEE SBAC-PAD*, pages 1–10, Bordeaux. IEEE Computer Society.

[Oda et al. 2018] Oda, R., Cordeiro, D., and Braghetto, K. R. (2018). Dynamic resource provisioning for scientific workflow executions in clouds. In *SCC*, pages 291–294. IEEE.

[Rocha et al. 2022] Rocha, F. W., Fukuda, J. C., Francesquini, E., and Cordeiro, D. (2022). Accelerating smart city simulations. In *Latin American High Performance Computing Conference*, pages 148–162. Springer.

[Vasconcelos et al. 2022] Vasconcelos, M. F. S., Cordeiro, D., and Dufossé, F. (2022). Indirect network impact on the energy consumption in multi-clouds for follow-the-renewables approaches. In *11th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS 2022)*, pages 44–55. SciTePress.