Harnessing solar energy via globally-distributed data centers

Abstract

Data centers have begun adapting to renewable energy sources like wind and solar in recent years to reduce operational carbon emissions. As the percentage of renewables increases in the electricity grid, time-shifting and space-shifting techniques have been employed to *gradually* align data center power usage with variable availability of renewable energy.

In this work, we propose a future computing architecture where data centers around the globe rely primarily on solar energy. To cope with the variation of solar power, we move workloads away from locations at night and instead run them at locations in daytime. We present the initial design of this architecture, and raise several important questions on how to transition from today's mixed-energy-powered data centers to future solar-powered data centers.

1 Introduction

Data center energy usage and carbon footprint have been exploding rapidly thanks to the increasing computing demands from artificial intelligence and machine learning, and the move to the cloud [1]. To reduce their operational carbon emissions, operators have begun exploring ways to better adopt low-carbon renewable energy sources, like solar and wind. This is a challenging problem because these renewables often have variable availability, and thus there are always times of day and locations where there's little renewable energy available.

To address this variable availability issues, several studies have looked into time-shifting and space-shifting techniques to align data center power usage with the availability of renewables [3, 5–7]. Because data centers today are powered by a mix of low-carbon renewables and high-carbon energy sources, these studies have been focused on how to *gradually* change existing data centers in response to growing renewables, but ultimately data centers need to be powered primarily by these time- and space-varying renewables.

In this paper, we explore a future computing architecture where a series of globally-distributed data centers that are powered primarily by solar. In order to reduce the use of stable power that's like high-carbon, we aggressively move workloads to locations with ample solar power to reduce. Because of the fairly predictable nature of solar power, we can move workloads to "chase" the sun, as shown in Figure 1. This can be achieved by reducing workloads at locations where the sun is about to set and starting new workloads at locations in their early morning hours.

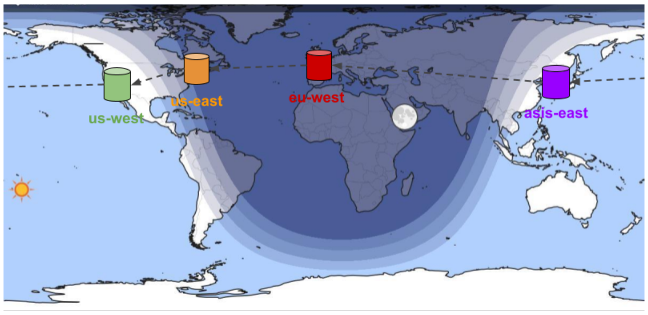

Figure 1. Coarse-grain solar energy availability in globally-distributed solar-powered data centers

We recognize that this design creates additional challenges in workload scheduling and data center capacity, but future renewable-based society may require more computing to be powered by these variable renewables.

Thus in this work, we want to quantitatively evaluate the cost of this solar-powered data center architecture, and compare it with existing data centers that do not move workloads around aggressively. In particular, we emphasize on these questions:

- how to schedule workloads around these globally-distributed solar-powered data centers?
- what's the impact of moving to this renewable-based architecture for each type of workload?
- what's the additional hardware cost to support varying amounts of workloads in this architecture?
- what's the minimum stable power required to support workloads that cannot be moved?

2 Globally-distributed Solar-powered Data Centers

Our proposed architecture consists of a series of globally-distributed data centers across multiple time zones, as shown in Figure 1. This figure shows the daytime and nighttime on a global scale, which is a coarse-grain approximation of solar energy availability. A few data centers across different time zones are shown here, which represents several major public cloud locations.

Note that at any given time, there will always be roughly half of the world lit by the sun, and thus up to half of the data centers can directly use solar energy. Our goal is to run most workloads at these locations to maximize solar energy usage and move workloads away from locations at their nighttime. As the earth rotates and the daytime region shifts westbound, we can move workloads along with the sun.

In this particular case, we can see that `us-west` and `asia-east` are in daytime and `us-east` and `eu-west` are in nighttime. Thus it's desired to run more workloads in these locations now. But as time goes by, we can see that `us-west` will start to lose solar power as the sun sets. Thus we need to stop assigning new workloads or stop existing workloads at `us-west`, and instead, move or run new workloads at data centers to its west, as shown in the arrows in Figure 1.

3 Challenges

Although solar-powered data centers can greatly reduce its operational carbon footprints, this new architecture poses several major challenges to how we run data centers:

3.1 Scheduling challenges

Careful scheduling of workloads across data centers is needed. Given the distribution of data centers around the globe, their solar availability may vary and thus we need to schedule workload movement so that each data center "hands off" its work when it starts to lose solar power. We also need to consider the maximum capacity of each data center, the available WAN bandwidth between data centers and the time needed to complete the hand-off (section 3.2).

Not all workloads can be moved easily. Existing commercial time-shifting solutions are often limited to internal flexible workloads [6] to avoid SLO violation, and space-shifting solutions are limited to short distances [3], due to their high impact on latency [4]. However, in future solar-powered data centers, such rigid workloads can incur too much carbon emission. Thus it is desirable to move as many workloads to "follow the sun" as possible, and we need to quantitatively measure the impact of moving each type of workload to determine the feasibility. This includes the overhead of setting up a new runtime, duplicating the input data, copying input/output data among data centers, and for online services, higher latency and WAN usage.

3.2 Capacity challenges

Additional WAN capacity is needed to absorb migration traffic. As shown in [3], even moving VMs between three data centers within a small region can result in up to 40% of existing WAN capacity. To support continuous workload movement around the globe, we need to measure the additional WAN bandwidth needed to move these workloads, and apply techniques like pause-and-restart [7], careful selection of replication sites and planned replication ahead of time. Furthermore, this can impact workload scheduling decisions, as workloads depending on large amounts of data need to wait longer for data transfers to start running.

Extra hardware capacity is needed to absorb larger power variations of renewable energy sources. Due to the variation of available renewable power, future data centers will likely have higher fluctuation in loads across times of day [2]. This will be especially true for solar-powered data centers where there's a clear daily pattern. To accommodate such power fluctuation, operators can either increase processor frequency (at higher energy costs, but less of an issue when using abundant solar energy) or purchase additional hardware, which will increase the embodied carbon footprint of data centers [2]. Thus, careful consideration must be taken to balance the reduction in operational carbon emissions and the increase in embodied carbon footprint, and it's important to determine the scale of power fluctuation in this solar-powered data center model.

4 Conclusion

In this work, we explore a future computing architecture where data centers around the globe are primarily powered by solar, and we shift workloads to "chase" the sun to reduce the use of other high-carbon energy sources.

Although this design may introduce additional scheduling and capacity challenges, it is a necessary step to reduce the operational carbon footprint of data centers. We lay out the major challenges to transition today's data centers to this future architecture, and plan to incorporate further quantitative analysis to determine the cost and feasibility of such approach.

References

[1] 2021. TechBriefs computing and Climate Change - Acm.org. https://www.acm.org/binaries/content/assets/public-policy/techbriefs/computing-and-climate-change-nov-2021.pdf

[2] Bilge Acun, Benjamin Lee, Kiwan Maeng, Manoj Chakkaravarthy, Udit Gupta, David Brooks, and Carole-Jean Wu. 2022. A Holistic Approach for Designing Carbon Aware Datacenters. *arXiv preprint arXiv:2201.10036* (2022).

[3] Anup Agarwal, Jinghan Sun, Shadi Noghabi, Srinivasan Iyengar, Anirudh Badam, Ranveer Chandra, Srinivasan Seshan, and Shivkumar Kalyanaraman. 2021. Redesigning Data Centers for Renewable Energy. In *HotNets*. https://www.microsoft.com/en-us/research/publication/redesigning-data-centers-for-renewable-energy/

[4] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. 2012. It's Not Easy Being Green. *SIGCOMM Comput. Commun. Rev.* 42, 4 (aug 2012), 211–222. https://doi.org/10.1145/2377677.2377719

[5] Íñigo Goiri, Kien Le, Thu D. Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. 2012. GreenHadoop: Leveraging Green Energy in Data-Processing Frameworks. In *Proceedings of the 7th ACM European Conference on Computer Systems* (Bern, Switzerland) *(EuroSys '12)*. Association for Computing Machinery, New York, NY, USA, 57–70. https://doi.org/10.1145/2168836.2168843

[6] Ana Radovanovic, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, et al. 2021. Carbon-aware computing for datacenters. *arXiv preprint arXiv:2106.11750* (2021).

[7] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In *Proceedings of the 22nd International Middleware Conference* (Québec city, Canada) *(Middleware '21)*. Association for Computing Machinery, New York, NY, USA, 260–272. https://doi.org/10.1145/3464298.3493399