Datacenter Demand Response for Carbon Mitigation: From Concept to Practicality

(Invited Paper)

Jiali Xing
University of Pennsylvania, Philadelphia, PA, USA
xjiali@seas.upenn.edu

Benjamin C. Lee
University of Pennsylvania, Philadelphia, PA, USA
leebcc@seas.upenn.edu

***Abstract*—Datacenters could greatly improve their sustainability by shifting their power usage across time in response to electricity's carbon intensity. However, despite decades of research, modulating demand in response to grid signals has not been deployed in wide practice. We review diverse studies and real-world practices to understand the reasons for the gap between concept and reality, identifying several significant challenges. Demand response frameworks (i) are often too complex, (ii) break abstraction layers in datacenter management, (iii) place too much emphasis on batch processing jobs, (iv) lack dynamic strategies, and (v) provide insufficient incentives for datacenter operators and users. Overcoming these challenges is essential to making datacenter demand response a reality, which would lead to sustainable, efficient computing.**

***Index Terms*—sustainability, datacenters, renewable energy**

I. Demand Response and Carbon

Hyperscale datacenters consume tens of terawatt hours of energy each year even as they continue to invest in new computational capacity. For instance, technology companies doubled their energy use between 2019 and 2022 [1], [2]. This growth may accelerate with the advent of emerging workloads such as artificial intelligence. Technology companies are increasingly motivated to manage their electricity usage better and mitigate their datacenters' operational carbon footprints. They invest heavily in renewable energy generation to offset datacenter consumption [3], [4]. However, reliance on intermittent wind and solar assets makes it increasingly difficult to match their computational demands with carbon-free electricity supplies [5], [6].

Carbon-conscious demand response (CDR) adjusts computational demand in response to variations in carbon-free energy supply, aligning datacenter power use to periods when electricity's carbon intensity is lower. Some researchers directly schedule batch jobs according to carbon signals, usually with greedy heuristics or mixed-integer programming [7]–[13]. Others predict and plan workload power usage [14], [15] or battery charge/discharge [16], [17]. Markets and game theory can apportion power adjustments and reduce carbon footprints [18]–[22]. Indeed, demand response and managing datacenters to exploit carbon-efficient electricity has been a broad topic of research that spans several decades [23].

Acknowledgements: National Science Foundation grant CCF-2326606 (Expedition in Computing, Carbon-Connect)

CDR has sought to mitigate carbon at multiple scales. Individual cloud users could modulate their workload [11]–[13]. Datacenters of various sizes could do so based on their installations and purchases of carbon-free energy [7], [9], [15], [20], [21]. And finally, micro- and power grids could do so by engaging with datacenters that are often among the largest loads in the system [14], [17], [18]. A system's scale and boundaries will often dictate the type of technical solution used for CDR. Direct scheduling can be effective for cloud users or privately owned datacenters because workloads are scheduled by a centralized orchestration framework. However, indirect incentives may be required for more complex environments—multi-tenant datacenters, hyperscale datacenters, or power grids—that are shared by diverse stakeholders.

Despite extensive research literature and recent industry reports, CDR has only been a concept rather than a practice [1], [2], [4]. Researchers and grid operators often identify datacenters as ideal candidates for CDR programs due to their large, supposedly flexible load [24], [25]. But datacenter operators themselves are often reluctant to participate in these programs and, to our knowledge, CDR is not currently deployed at scale [1], [2], [4], [26]. Only in 2022 did Google implement and test a CDR framework, but this solution was designed for grid stability and power emergencies [15], [27].

II. Challenges and Opportunities

Model Complexity. Existing literature often describes complex, end-to-end optimization frameworks that seek comprehensive solutions for carbon mitigation [10], [14], [17], [28]–[30]. While theoretically optimal, these models are intricate and fragile, making them impractical for production systems. They require extensive modeling process and precise centralized control, which can be difficult in dynamic, distributed systems. Furthermore, prior research focuses on sophisticated optimization but makes simplistic assumptions about datacenter operations. For example, they often assume a large percentage of workload power can be flexibly rescheduled without significant performance loss or difficulty.

We need more practical models for real-world systems. Workload analysis will enable schedules that better balance performance with carbon reductions. Performance and power models will need to better predict the impact on energy

consumption and operational efficiency. Furthermore, understanding the actual flexibility of workloads across time periods or geographical locations is critical. We need to formulate and satisfy constraints imposed by various workloads as we exploit temporal and spatial variations in electricity's carbon intensity. Machine learning and real-time data analytics could enable models that better capture variability.

Abstractions and Interfaces. Practical CDR must account for robust abstractions and interfaces that have enabled modern datacenter scaling. CDR must interface with grids that supply power through sophisticated purchase agreements and pricing mechanisms [18]. It must simultaneously interface with loosely federated schedulers that manage diverse workloads [31]. Previous proposals neglect these layers of abstraction and directly schedule jobs based on the grid's carbon signals [9], [10]. Yet frameworks that allocate power without accounting for job scheduling might over-estimate workload flexibility and under-estimate performance losses [15], [18], [32].

Future research must re-think abstractions and interfaces between CDR frameworks, power infrastructure, and workload management in datacenters. Some of these interfaces may require scenario planning because there exists uncertainty about the role of power purchase agreements and renewable energy credits when assessing the net carbon intensity of a datacenter's electricity. Location-based estimates analyze the local grid's energy sources. Market-based estimates account for renewable energy certificates generated by datacenter operators' investments in wind and solar projects. How carbon is assessed and communicated will impact CDR policy.

We draw inspiration from successes such as Google's Carbon-Intelligent Computing System [15], [33], which integrates CDR with the Borg scheduler to preserve existing abstractions. When CDR curtails or boosts power, the number of virtual machines available for Borg allocation decreases or increases, respectively. CDR inherits existing scoring functions that govern how jobs are assigned to machines; there would simply be more or fewer machines depending on CDR policy.

Workload Diversity. Prior research has disproportionately focused on scheduling batch (*i.e.* offline, background) computation. Such jobs are assumed to be deferrable, permitting simpler workload models and optimization. However, batch workloads are not as flexible as typically assumed [31]. They are often launched with specific landing times (*e.g.*, four hours) and cannot be deferred arbitrarily. Moreover, interactive (*i.e.* online, real-time) workloads account for a large portion of datacenter power [31]. These workloads are less amenable to CDR policies designed to target batch workloads. CDR strategies that reduce carbon through batch workloads alone risks prohibitively large penalties for these workloads.

Future CDR should accommodate a diverse array of batch and interactive workloads. Techniques should pursue efficiency, aligning power curtailments and boosts to workloads and computational phases that suffer least and benefit most, respectively. Techniques should also balance efficiency with fairness, distributing power adjustments across workload types so that all users' workloads bear some responsibility for carbon mitigation. Finally, we could curate workload libraries in which a workload specifies multiple versions, each with distinct performance and power characteristics, allowing CDR frameworks to dynamically switch between versions based on energy availability or carbon intensity. More granular workload options permit more precise trade-offs between service-level objectives and carbon intensity.

Dynamic Strategies. Day-ahead forecasts for computational load and carbon intensity are increasingly accurate due to recent advances in machine learning [15], [32], [34], [35]. These forecasts often provide fixed inputs to optimizations, such as linear programming, that solve for the next day's hourly power allocations. But such allocations cannot flexibly respond to real-time variations in datacenter or grid conditions.

Future approaches to CDR could include dynamic mechanism design, a game-theoretic framework that makes decisions in an environment where the game evolves over time and where agents' preferences, information, and constraints evolve over time. Unlike static mechanisms, which assume a single one-shot interaction between agents, dynamic mechanisms anticipate repeated interactions and accommodates uncertainty about future events. Multi-agent reinforcement learning could prove helpful [36], [37], but research is required to guarantee grid stability and datacenter availability.

Incentive Structures. Most existing approaches to CDR do not incentivize datacenters and their users to modulate power usage. Datacenter users seek performance whereas datacenter operators seek energy efficiency and, increasingly, carbon reductions. Techniques that improve efficiency often offer benefits to operators while only offer risks to users, leading to misaligned incentives. The absence of carbon attribution, which estimates each job's contribution towards the system's total carbon footprint, prevents effective decision making [38].

We need carbon attribution techniques that are transparent, fair, and fine-grained. Operational carbon will depend on power use and the grid's carbon intensity whereas embodied carbon will depend on hardware use and the amortization of manufacturing costs across time. Attribution permits carbon pricing, which could be implemented by cloud providers independently of any broader regulatory policy.

Multi-agent game theory provides a framework in which users and jobs act selfishly to pursue performance objectives within a game that is designed to achieve broader system objectives for carbon efficiency. An effectively designed game would allow strategic users to engage with CDR on their own terms, determining how much power to use and when. Under what conditions might a user forgo 10KW now for 15KW later? Users (or intelligent agents that act on their behalf) could learn and optimize policies that respond to such choices. And datacenter operators could, in turn, explore and optimize the set of choices offered to produce desired system outcomes.

III. Conclusion

While demand response offers significant promise for reducing datacenter computing's carbon footprint, several challenges remain. Existing approaches have defined the solution space

and demonstrated potential, but we require further research that extends demand response to more realistic models of computational load and system interfaces. Moreover, further research must explore dynamic strategies that incentivize participation in demand response and contributions to carbon mitigation. Whereas mandates enforced by centralized schedulers could prove effective, we find incentives that motivate decentralized decision making more attractive given the diversity of stakeholders in modern hyperscale datacenters.

REFERENCES

[1] Google, "2024 Environmental Report."

[2] Meta, "2023 Sustainability Report."

[3] Amazon, "Amazon Sustainability | 2023 Report."

[4] Microsoft, "2024 Environmental Sustainability Report | Microsoft CSR."

[5] C. I. S. Operator, "California ISO - Managing Oversupply."

[6] A. Bunodiere and H. S. Lee, "Renewable Energy Curtailment: Prediction Using a Logic-Based Forecasting Method and Mitigation Measures in Kyushu, Japan," *Energies*, vol. 13, no. 18, Jan. 2020.

[7] . Goiri, K. Le, M. E. Haque, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "GreenSlot: scheduling energy consumption in green datacenters," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. Association for Computing Machinery, 2011.

[8] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1, 2012.

[9] . Goiri, W. Katsak, K. Le, T. D. Nguyen, and R. Bianchini, "Parasol and GreenSwitch: managing datacenters powered by renewable energy," in *Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems*. Association for Computing Machinery, Mar. 2013.

[10] Y. Zhang, D. C. Wilson, I. C. Paschalidis, and A. K. Coskun, "HPC Data Center Participation in Demand Response: An Adaptive Policy With QoS Assurance," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 1, Jan. 2022.

[11] W. A. Hanafy, Q. Liang, N. Bashir, D. Irwin, and P. Shenoy, "CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing Carbon-Efficiency," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 7, no. 3, Dec. 2023.

[12] A. Lechowicz, N. Christianson, J. Zuo, N. Bashir, M. Hajiesmaili, A. Wierman, and P. Shenoy, "The Online Pause and Resume Problem: Optimal Algorithms and An Application to Carbon-Aware Load Shifting," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 7, no. 3, Dec. 2023.

[13] W. A. Hanafy, Q. Liang, N. Bashir, A. Souza, D. Irwin, and P. Shenoy, "Going Green for Less Green: Optimizing the Cost of Reducing Cloud Carbon Emissions," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*. Association for Computing Machinery, Apr. 2024.

[14] L. Yu, T. Jiang, and Y. Zou, "Distributed Real-Time Energy Management in Data Center Microgrids," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, Jul. 2018.

[15] A. Radovanovic, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care, S. Talukdar, E. Mullen, K. Smith, M. Cottman, and W. Cirne, "Carbon-Aware Computing for Datacenters," *arXiv:2106.11750 [cs, eess]*, Jun. 2021.

[16] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and limitations of tapping into stored energy for datacenters," in *Proceedings of the 38th annual international symposium on Computer architecture*. Association for Computing Machinery, Jun. 2011.

[17] B. Wang, C. Zhang, and Z. Y. Dong, "Interval Optimization Based Coordination of Demand Response and Battery Energy Storage System Considering SOC Management in a Microgrid," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, Oct. 2020.

[18] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," in *The 2014 ACM international conference on Measurement and modeling of computer systems*. Association for Computing Machinery, Jun. 2014.

[19] M. A. Islam, H. Mahmud, S. Ren, and X. Wang, "Paying to save: Reducing cost of colocation data center via rewards," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, Feb. 2015.

[20] M. A. Islam, X. Ren, S. Ren, A. Wierman, and X. Wang, "A market approach for handling power emergencies in multi-tenant data center," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Mar. 2016.

[21] M. Islam, X. Ren, S. Ren, and A. Wierman, "A Spot Capacity Market to Increase Power Infrastructure Utilization in Multi-tenant Data Centers," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Feb. 2018.

[22] M. R. Hossen, K. Ahmed, and M. A. Islam, "Market Mechanism-Based User-in-the-Loop Scalable Power Oversubscription for HPC Systems," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Feb. 2023.

[23] F. Kong and X. Liu, "A Survey on Green-Energy-Aware Power Management for Datacenters," *ACM Computing Surveys*, vol. 47, no. 2, Nov. 2014.

[24] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response," in *International Green Computing Conference*. IEEE, Nov. 2014.

[25] M. Chen, C. Gao, M. Song, S. Chen, D. Li, and Q. Liu, "Internet data centers participating in demand response: A comprehensive review," *Renewable and Sustainable Energy Reviews*, vol. 117, Jan. 2020.

[26] B. Acun, B. Lee, K. Maeng, M. Chakkaravarthy, U. Gupta, D. Brooks, and C.-J. Wu, "A Holistic Approach for Designing Carbon Aware Datacenters," *arXiv:2201.10036 [cs, eess]*, Jan. 2022.

[27] V. Mehra and R. Hasegawa, "Supporting power grids with demand response at google data centers."

[28] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. Association for Computing Machinery, 2011.

[29] M. Zeraati, M. E. Hamedani Golshan, and J. M. Guerrero, "Distributed Control of Battery Energy Storage Systems for Voltage Regulation in Distribution Networks With High PV Penetration," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, Jul. 2018.

[30] W. Deng, F. Liu, H. Jin, C. Wu, and X. Liu, "MultiGreen: cost-minimizing multi-source datacenter power supply with online control," in *Proceedings of the fourth international conference on Future energy systems*. Association for Computing Machinery, May 2013.

[31] B. Acun, B. Lee, F. Kazhamiaka, A. Sundarrajan, K. Maeng, M. Chakkaravarthy, D. Brooks, and C.-J. Wu, "Carbon Dependencies in Datacenter Design and Management," in *HotCarbon'22*, 2022.

[32] B. Acun, B. Lee, F. Kazhamiaka, K. Maeng, U. Gupta, M. Chakkaravarthy, D. Brooks, and C. Wu, "Carbon explorer: A holistic framework for designing carbon aware datacenters," in *Proc. Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2023.

[33] M. Tirmazi, A. Barker, N. Deng, M. E. Haque, Z. G. Qin, S. Hand, M. Harchol-Balter, and J. Wilkes, "Borg: the next generation," in *Proceedings of the Fifteenth European Conference on Computer Systems*. Association for Computing Machinery, 2020.

[34] D. Maji, P. Shenoy, and R. K. Sitaraman, "CarbonCast: multi-day forecasting of grid carbon intensity," in *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. Association for Computing Machinery, Dec. 2022.

[35] D. Maji, R. K. Sitaraman, and P. Shenoy, "DACF: day-ahead carbon intensity forecasting of power grids using machine learning," in *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. Association for Computing Machinery, Jun. 2022.

[36] S. Fan, S. Zahedi, and B. Lee, "The computational sprinting game," in *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2016.

[37] C. Yeh, V. Li, R. Datta, J. Arroyo, N. Christianson, C. Zhang, Y. Chen, M. M. Hosseini, A. Golmohammadi, Y. Shi *et al.*, "Sustaingym: Reinforcement learning environments for sustainable energy systems," *Advances in Neural Information Processing Systems*, 2024.

[38] L. Han, J. Kakadia, B. C. Lee, and U. Gupta, "Towards game-theoretic approaches to attributing carbon in cloud data centers," in *Proceedings of the 2024 HotCarbon Workshop*. ACM, 2024.