Carbon Dependencies in Datacenter Design and Management

Bilge Acun[1], *Benjamin Lee*[1,2], *Fiodar Kazhamiaka*[1,3], *Aditya Sundarrajan*[1], *Kiwan Maeng*[1], *Manoj Chakkaravarthy*[1], *David Brooks*[1,4], *Carole-Jean Wu*[1]

[1]*Meta,* [2]*University of Pennsylvania,* [3] *Stanford University,* [4]*Harvard University*

Abstract

Building a sustainable datacenter requires coordinated decisions in its design and system management. Existing research work on datacenter sustainability often considers the design space in isolation and misses opportunities to minimize carbon footprint through coordinated design and management where sustainability is a first-class objective. Design decisions such as datacenter site selection, renewable energy investment portfolios, and the provisioning of energy storage are intertwined with complementary solutions for operation, including various forms of demand response and carbon-aware workload management. In this paper, we advocate for holistic frameworks that take into account both *operational and embodied carbon* to coordinate between datacenter design and system management decisions.

1 Introduction

Information and Communication Technology (ICT) sector's carbon emissions are now estimated at 1.8-3.9% of global emissions [10], and continues to grow as the world becomes more digitized. Large technology companies are exploring ambitious sustainability strategies to be 24/7 carbon free in their operations [15, 22] or carbon-free in their supply chains [26], motivating us to rethink how datacenters are operated and designed[1].

The contemporary strategy of using carbon offsets or annual renewable energy credits (RECs) to achieve carbon-free operations [12, 21, 23] does not address the mismatch between renewable supply and demand on an hourly basis. As the share of renewable energy in the grid continues to grow, this mismatch can cause increasingly drastic curtailments (i.e. deliberate reduction in renewable output). To highlight this issue, in 2021 renewable energy curtailments reached 6% of the total generated renewable energy in the California grid, which has ≈ 33% share of renewables [6, 32]. These curtailments deactivate renewable energy generation to match supply with demand [4, 8, 24]. Given these conditions and the projected growth of the renewable energy generation [35] — the majority coming from wind and solar — it is clear that relying on annual renewable energy credits is not a sustainable long-term strategy. We believe that complementary solutions to renewable deployment, such as demand response and battery deployment, will play an increasingly important role in reducing the carbon footprint of datacenters.

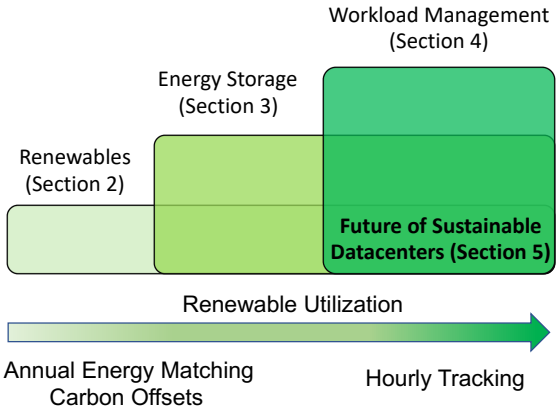

Figure 1: Sustainability solutions are interdependent with each other and a coordinated solution with fine grained measurements (hourly tracking) is necessary to reach a carbon-optimal solution and to improve renewable utilization.

This research argues for the adoption of integrated solutions to design and manage sustainable datacenters. Designing a *carbon-optimal* datacenter — one that minimizes its lifetime carbon footprint — requires us to navigate trade-offs and dependencies across the design and management decisions of power infrastructures, computational hardware, and workload requirement. For example, while the prime focus of building renewable infrastructures and its complementary solutions is to reduce **operational footprint** from the electricity use of datacenters, an important but under-explored aspect of the design space in exploring these strategies is **embodied**

[1]This work represents the research carried out at Meta's Fundamental AI Research (FAIR). The opinions represent our research organization's views and are framed in the context of the broader industry. It does not intend to reflect any of Meta's datacenter plans.

footprint — the carbon footprint coming from the manufacturing of the hardware. The majority of the existing work has focused on minimizing the operational footprint [2, 30, 36]; however, taking embodied footprint into account can significantly change the system design. When embodied carbon footprint is considered, carbon-optimal datacenters may ***not*** be able to run at 100% renewable coverage 24/7 because of the manufacturing footprint of renewable farms and batteries [1]. Similarly, workload management strategies that engage in demand-response may require additional server capacity to support the increased demand during peak renewable energy hours, creating a carbon trade-off with the embodied footprint of the servers.

Sustainability should not be approached in isolation for each datacenter layer — the dependencies created by sustainability strategies require a cross-layer co-design that treats sustainability objectives as a first-class principle. We need coordinated strategies for deploying renewable energy generation, energy storage, and demand response scheduling in datacenters across the globe. We also need holistic frameworks that allow us to adequately explore this space. As visualized in Figure 1, implementing these solutions in coordination can increase renewable energy utilization and enable a more carbon-efficient datacenter system. In this paper, we discuss the following aspects of the solution space:

- Renewables: Investment decisions, accounting mechanisms, and the interface between power grid and datacenters (§ 2)
- Energy storage: the role of batteries in datacenters and its challenges (§ 3)
- Workload management: understanding the workload characteristics at scale and opportunities for demand-response techniques (§ 4).
- Implications on future datacenter design: need for holistic frameworks to guide the datacenter design by coordinating strategies for deploying renewable energy, energy storage, and demand-response (§ 5).

2 Renewables and the Grid

Currently, datacenter (DC) operators invest in renewable generation, such as wind and solar, and implement power purchase agreements (PPAs) to reduce DC exposure to a grid's carbon intensity. PPAs link *Renewable Energy Credits (RECs)* with a specific source of energy and issue, e.g., one certificate for every MWh generated [11, 13, 22]. RECs are commonly issued on an annual basis. Hourly accounting of renewable energy, *Time-based Energy Attribute Certificates (T-EACs)*, is necessary for more fine-grained, accurate accounting. T-EACs are in development and need industry standardization [31].

Existing grid interfaces provide foundations for carbon accounting, which define a DC operator's progress toward green computing. PPAs that generate RECs allow operators to achieve its *Net Zero* objectives by ensuring RECs offset datacenter energy consumed at the year end. PPAs that generate hourly-RECs (or T-EACs) allow operators to go further and pursue 24/7 carbon-free objectives by ensuring RECs offset DC energy consumed in every hour of the year. However, using daily or hourly RECs does not mean renewable energy is directly used by the DC. RECs may be generated at a different location. Therefore, despite these specific advances in financial and accounting mechanisms, researchers should increasingly take a broader perspective on the interface between the DC and grid.

Local versus Global Optimization. Researchers should examine the effects of RECs and T-EACs on the broader energy grid. Current financial and accounting mechanisms provide DCs a narrow set of actions and strategies for locally optimizing computation for their own sustainability goals. But the DC's local optimum may differ, perhaps significantly, from the grid's global optimum. Suppose a DC seeks to minimize its operational carbon footprint when renewable energy is scarce. The DC defers jobs only to the extent needed to align its energy consumption with the RECs from its investments in renewable energy. It neglects opportunities to further identify and defer flexible jobs because doing so incurs a performance cost with no additional benefit to *Net Zero* or 24/7 carbon-free objectives. This strategy is optimal for the DC but sub-optimal for the grid, which seeks responsive and flexible loads that help align energy demand and supply throughout its network. In effect, today's DCs experience the benefits of sustainability using RECs while the grid incurs the risks of balancing energy demand and supply.

When renewable energy is abundant, the grid often suffers from insufficient load and curtails generation. Data from the California Independent System Operator (CAISO) suggests that curtailments increase in frequency and magnitude as the number of solar and wind farms increases [5]. Avoiding curtailments requires energy storage and/or shifting more energy consumption to renewable generation peaks, which in turn requires more effective time-of-use pricing which encourages the grid's largest consumers (e.g., DCs) to shape their demands.

Datacenter Site Selection. Operators must decide where to site their datacenters and, by implication, decide which grids should supply energy. These decisions impact carbon footprints in several dimensions. First, in the near and medium term, some datacenters may continue to use the broad grid's energy mix and the carbon-intensity of the grid can vary significantly by geography. Second, operators might mitigate risks by investing in a diversified mix of renewable energy types (e.g., wind, solar, geothermal) and some geographic locations may offer broader and consistent supplies. Finally, operators might mitigate risks by shifting computational load between geographic locations. To further this goal, datacenters should reside in complementary locations with uncorrelated peaks and valleys in renewable generation.

Selecting a site to set up or lease datacenter space is an important decision that last for several years or even decades. Beyond sustainability, there are many factors that affect DC site selection, including the availability of land and labor, reliable water and power supply, geographic fault tolerance (in the case of multi-datacenter deployments) and proximity to end users. Research is needed to reconcile these classic constraints with emerging ones that impact sustainability.

3 Energy Storage

Batteries are expensive [9] and have an embodied carbon footprint that is amortized over limited calendar lifetimes. Their operational characteristics may degrade over time and after numerous (dis)charge cycles. These properties should be part of the calculus for carbon-aware datacenter design and management and be weighed against the many benefits that they provide. Batteries have many applications in datacenters – several of which we discuss next – and these applications compete for battery resources and create dependencies between datacenter design and management decisions.

3.1 Battery Applications in Datacenters

Renewable Energy Buffer. A battery can reduce a datacenter's reliance on dispatchable, carbon-intensive power sources while increasing the use of intermittent sources like wind and solar by buffering energy when there is a generation surplus. This could allow the datacenter operator to reach their target carbon footprint with smaller investments in wind and solar farms or with less disruptive demand response scheduling. A large datacenter-scale battery could provide these services to the regional electricity network.

Ancillary Grid Services. A battery could provide reliability and stability services to the grid, including frequency regulation, ramping, and others [7, 28]. The decision to provide these services at any given moment would depend on the anticipated load on the battery, and should be made in conjunction with workload management decisions. Datacenter operators may require dynamic prices and incentives when storing and supplying renewable energy for the grid during periods of abundance and scarcity, respectively.

Uninterruptible Power Supply (UPS). Storage for renewable energy is an extension of existing datacenter power infrastructure. Batteries are often used in datacenters as a back-up power source for server racks that last up to tens of minutes until back-up generators come online [25]. UPS's are idle the vast majority of the time and can be replaced by distributed battery modules that collectively function as one large battery for demand response applications while reserving some energy to ensure resilience and availability for their local rack [29]. Larger, utility-scale batteries might also provide an alternative to carbon-intensive back-up generators.

3.2 Battery Dependencies

The questions of battery design and management are dependent on each other, and tied to its many applications and design/management decision of other parts of the DC. To decide how to size the battery, one must consider how the applications will compete to charge, discharge, or reserve the storage capacity. Similarly, battery management decisions will depend on the size and placement of the battery. Batteries can be placed on the DC site or at the utility/grid in close proximity of the renewable farms. Placement on-site enables DC operators to exert more control over battery charge/discharge decisions. On the other hand, placement on the grid has a number of broader advantages. By investing in batteries, DC operators can help match supply and demand for the broader grid and reduce the transmission line bottlenecks by placing the batteries close to the energy source. Thus, the relative merits of on-grid and on-site battery requires further studies.

Battery design and management decisions should also be coordinated with workload management. More flexibility in the workload can help reduce the total battery capacity required to reach a target carbon footprint, and vice versa.

4 Workload Management at Scale

Datacenter demand response requires a better understanding of the workload characteristics at scale (4.1) and re-thinking workload management(4.2).

4.1 Workload Characteristics

Hyperscale datacenters such as those run by Google, Microsoft, Meta, Amazon are globally distributed and typically run a mix of delay-sensitive and delay-tolerant workloads. Delay-sensitive workloads such as real-time user facing requests have strict service level objectives (SLOs) for execution and completion times. Therefore, they cannot be shifted in time but can be shifting in space (i.e. to a different datacenter location with lower carbon intensity). Delay-tolerant workloads, on the other hand, have less stringent SLOs and can be shifted both in space and time and executed when the necessary resources are available at lower carbon intensity. Examples of delay-tolerant workloads include offline machine learning training jobs, daily data processing jobs, etc.

Temporal and Spatial Flexibility. In large datacenter deployments, delay-tolerant workloads constitute a large portion of the total datacenter power demand. Delay-tolerant workloads have varying SLOs depending on how important their execution is to other dependencies. For instance, Google has reported that flexible jobs with 24-hour completion SLOs make up about 40% of the Borg scheduler's jobs [33]. At Meta, around 20-30% of all workloads are delay-tolerant with varying SLOs. These constitute a mix of offline data processing, offline training and opportunistic compute workloads. As an example of the SLO breakdown, in Figure 2, we plot the breakdown of data processing workloads by completion time

SLOs. The data processing workloads constitute about 7.5% of all the workloads in the fleet. Of these, about 87.4% of the workloads have SLOs that are greater than 4-hours with a majority having 24-hour SLOs. This provides great flexibility in workload time shifting to optimize carbon.

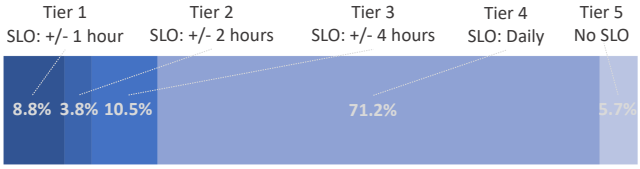

Figure 2: Breakdown of data processing workloads by completion time SLO at Meta.

Delay-tolerant workloads vary in the flexibility they provide for space and time shifting. For example, opportunistic state-less compute workloads are extremely flexible and can be scheduled in any DC with available resources and at any time when the least amount of carbon is used. On the other hand, offline training jobs are stateful, with data dependencies that need to be met, and may require specific hardware to run on. Hence, while these offline training jobs can be scheduled at a time that minimizes carbon [20], the requirement for special hardware and data dependencies could impose spatial locality restrictions for the scheduler.

In Figure 3, we characterize some of the Meta workloads based on time and space shifting flexibility. We classify workloads as less flexible in space if the workloads can only be scheduled in specific DC locations because of data or hardware dependencies. On the other hand, workloads are more flexible in space if they can be scheduled anywhere. Similarly, workloads are less flexible in time if they need to be scheduled and completed within a small time window and more flexible if they can be scheduled and completed within a larger time window. As an example, data processing workloads are flexible in time depending on the completion time SLOs as described in Figure 2 but are restricted in where they can run due to the presence of data dependencies. On the other hand, offline state-less compute workloads have similar time flexibility characteristic and, at the same time, are also flexible in where they are scheduled. In contrast, real time services such as the web services handling end-user requests have strict completion time SLOs and are somewhat restrictive on where they are scheduled to minimize end-user latency. Similarly, AI inference workloads which are used for real time recommendations have similar time inflexibility characteristics [17, 19] and are also less flexible in space due to hardware and data restrictions.

In general, restrictions around spatial flexibility can be addressed by distributing servers of different types more evenly across all datacenters, or by replicating training data in more datacenters to enable training everywhere. However, these options come at the cost of increased carbon usage due to this

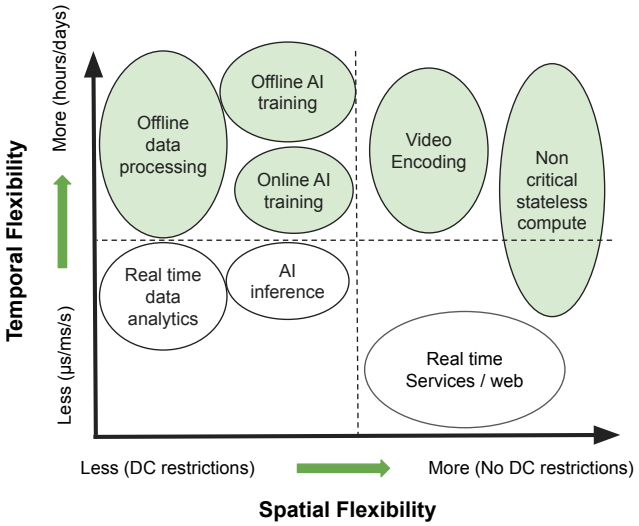

Figure 3: Characterizing the space and time flexibility of workloads that run in production datacenters. Delay-tolerant workloads are highlighted in green.

replication. Hence, while the presence of delay-tolerant workloads can provide opportunities to minimize carbon usage, we need to take a holistic approach by considering the time and space flexibility characteristics of these workloads and the added embodied and operational carbon cost of leveraging this flexibility.

Pre-computation for real time workloads. Power-intensive computations and data movement can also be performed in advance during the time when renewables are abundant. Real-time deep learning inference use cases at the cloud are supported by backend computation kernels that do not necessarily come with real-time constraints and can be processed offline ahead of time. Future systems must adopt code modularity by allowing kernels of a program to operate independently. By doing so, the finer-granularity of the code structure will allow computations to be scheduled, depending on renewable availability.

4.2 Workload Management Techniques

Datacenter demand response requires re-thinking resource management. Software services and jobs should receive larger resource allocations when renewable energy is abundant and receive smaller allocations when it is scarce. Realizing this goal requires shaping computational demand to create valleys and peaks that align with carbon-free energy supply. Supply depends on investments in renewable energy generation, purchasing strategies for that energy, and investments in energy storage. Demand depends on workload flexibility and price elasticity, which quantifies the change in resource consumption given a change in resource price.

Demand response differs significantly from current best practice in hyperscale datacenters. It has been shown that infrastructure costs associated with power delivery can be

amortized when the use of a datacenter's provisioned power is maximized. This strategy is implemented in two ways. First, batch jobs are scheduled on servers that are under-utilized by interactive services. Second, the datacenter deploys more servers than it can power simultaneously, relying on statistical load variations across complementary jobs and power capping to avoid oversubscribing the shared power supply. Taken together, these strategies eliminate the classic assumption of diurnal patterns in datacenter computing such that a datacenter built to support 30MW of computation consistently uses 30MW of power.

A demand response framework's first task is to recreate diurnal rhythms in computational load and energy demand in ways that align with time-varying supplies of renewable energy. Optimization provides one possible solution, [30] minimizing the difference between the datacenter's energy demand and the grid's renewable energy supply. The optimization would be constrained by datacenter's provisioned power, provisioned servers, and workload flexibility. The optimization would provide day-ahead schedules, determining each job's hourly resource allocation for the next twenty-four hours based on forecasts of energy demand and supply. Despite initial progress, however, forecasting and optimization may prove insufficient and we envision several significant, new research directions.

Coordinating Management. Datacenters require a coherent management strategy for servers and batteries. Coordination is required because servers and batteries compete for renewable energy. Servers seek energy for jobs immediately whereas batteries do so to accumulate energy for future computation. Batteries may extend the optimization objective by introducing new constraints based on how the number of (dis)charge cycles affect their lifetimes. Batteries will also make scalable optimization more difficult by increasing the number of variables to be optimized.

Scaling Management. Scalable optimization for datacenter scheduling may prove to be difficult. One formulation might require optimizing a variable for every hour in the day and for every server (or indeed every processor or core). Another formulation might require optimizing a variable for every job or task. Research will be required to formulate computationally tractable variants of the optimization problem, balancing the resource and time granularity of scheduling decisions with the solver's computational costs.

The costs of centralized optimization might motivate alternative decision making frameworks such as distributed, multi-agent systems. Agents could represent users and their jobs, developing and optimizing independent strategies for requesting server and power allocations based on workload needs, datacenter conditions (e.g., battery levels), grid signals (e.g., renewable energy supplies, prices), and expected competition between agents. Game theory could model system dynamics when agents independently pursue performance yet account for carbon costs. Mechanism design could structure the rules of the allocation game so that independent agents can act strategically yet produce a datacenter-wide equilibrium with desirable performance and sustainability outcomes.

Exploring Carbon Trade-Offs. Demand response produces interesting trade-offs between a DC's operational and embodied carbon footprints. Scheduling jobs to create diurnal loads that match renewable energy supply will reduce operational carbon. But diurnal patterns may produce peak loads that are much higher than today's expected loads. Serving these peaks may require many more servers than what today's datacenters have deployed and may increase embodied carbon [16,18]. Because these additional servers are utilized only occasionally during peak loads, these embodied carbon costs may not be fully amortized. Research is needed to balance demand response against its hardware requirements and, more broadly, to examine the role of dark silicon in sustainable datacenter computing [3].

Demand response also produces interesting trade-offs between power usage effectiveness (PUE) and carbon costs. Traditionally, hyperscale datacenters seek the lowest possible PUE (e.g., Meta's PUE of 1.1 [12], Google's of 1.1 [14], and Microsoft's of 1.18 [27]). When renewable energy is abundant, the datacenter might activate more machines, dissipate more power, and generate more heat. Under these circumstances, renewable energy may permit the adoption of more advanced cooling strategies that are not the most power-efficient but allow the datacenter to increase its computational throughput and serve peak loads yet remain carbon neutral.

5 Future Datacenter Design and Management

Building a sustainable DC that effectively utilizes renewable energy involves making decisions that affects both the design and management of the DCs. In addition, there needs to be a coordination between these decisions, as presented in this research. For example, on the design side, determining the renewable investment amounts or battery deployment capacity requires deep understanding of the energy generation characteristics of a region. Renewable characteristics affect DC site selection decisions. Similarly, workload shifting can reduce the amount of energy storage capacity needed. On the management side, DC-side battery deployment requires coordination between charge/discharge decisions and workload scheduling decisions, such as time/space shifting.

Another category of coordination is required to manage the carbon trade-off between operational and embodied footprint. For example, workload shifting may require additional server capacity to be built in DCs to allow room for scheduling more workloads to peak renewable energy hours. Building additional servers comes with embodied footprint.

Figure 4 presents the process of identifying a carbon-optimal DC design. A holistic design for a DC must consider both operational and embodied carbon when minimizing the overall carbon footprint. Operational inputs include hourly

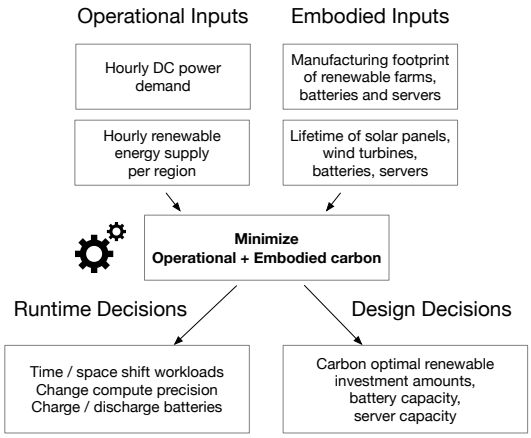

Figure 4: A framework for carbon-free datacenter design.

DC power demand and renewable power supply for the corresponding DC location. Embodied inputs account for the carbon emissions from manufacturing and the expected lifetimes of solar and wind farms, lithium-ion batteries, and datacenter servers. Finally, the outputs of the framework include both runtime and design decisions.

As a world, we have a limited carbon budget (230-440bn tonnes of CO_2($GtCO_2$) from 2020 onwards) before reaching Paris Agreement's $1.5°C$ target [34]. Our focus should be spending this budget onto solutions that maximize carbon reduction — in other words, minimizing the sum of operational and embodied carbon. This requires a holistic view of the datacenter design and management.

Acknowledgements

We would like to thank Kim Hazelwood for supporting this work and the anonymous reviewers for their feedback on this paper.

References

[1] Bilge Acun, Benjamin Lee, Kiwan Maeng, Manoj Chakkaravarthy, Udit Gupta, David Brooks, and Carole-Jean Wu. A holistic approach for designing carbon aware datacenters. *arXiv preprint arXiv:2201.10036*, 2022.

[2] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, 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*, pages 350–358, 2021.

[3] Erik Brunvand, Donald Kline, and Alex K. Jones. Dark silicon considered harmful: A case for truly green computing. In *2018 Ninth International Green and Sustainable Computing Conference (IGSC)*, pages 1–8, 2018.

[4] California ISO. Managing oversupply. http://www.caiso.com/informed/Pages/ManagingOversupply.aspx, 2021.

[5] California ISO. Managing oversupply, 2022.

[6] California State Commission. 2020 total system electric generation. https://www.energy.ca.gov/data-reports/energy-almanac/california-electricity-data/2020-total-system-electric-generation, 2020.

[7] Hao Chen, Zhenhua Liu, Ayse K Coskun, and Adam Wierman. Optimizing energy storage participation in emerging power markets. In *2015 Sixth International Green and Sustainable Computing Conference (IGSC)*, pages 1–6. IEEE, 2015.

[8] A. Chien. Characterizing opportunity power in the california independent system operator (caiso) in years 2015-2017. *Energy and Earth Science*, 2020.

[9] Wesley Cole, A Will Frazier, and Chad Augustine. Cost projections for utility-scale battery storage: 2021 update. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2021.

[10] ACM Technology Council. Computing and climate change. https://dl.acm.org/doi/pdf/10.1145/3483410, 2021.

[11] Facebook. Advancing renewable energy through green tariffs. https://sustainability.fb.com/wp-content/uploads/2020/12/FB_Green-Tariffs.pdf, 2021.

[12] Facebook. How we're helping fight climate change. https://about.fb.com/news/2021/06/2020-sustainability-report-how-were-helping-fight-cli 2021.

[13] Google. Google's green ppas: What, how and why. https://static.googleusercontent.com/media/www.google.com/en//green/pdfs/renewable-energy.pdf, 2013.

[14] Google. Google data center pue performance. https://www.google.com/about/datacenters/efficiency/, 2022.

[15] Google. Operating on 24/7 carbon-free energy by 2030. https://sustainability.google/progress/energy/, 2022.

[16] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. Act: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[17] Udit Gupta, Samuel Hsia, Vikram Saraph, Xiaodong Wang, Brandon Reagen, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. Deeprecsys: A system for optimizing end-to-end at-scale neural recommendation inference. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 982–995, 2020.

[18] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 854–867, 2021.

[19] Udit Gupta, Carole-Jean Wu, Xiaodong Wang, Maxim Naumov, Brandon Reagen, David Brooks, Bradford Cottel, Kim Hazelwood, Mark Hempstead, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, and Xuan Zhang. The architectural implications of facebook's dnn-based personalized recommendation. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 488–501, 2020.

[20] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, and Xiaodong Wang. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 620–629, 2018.

[21] Lucas Joppa. Made to measure: Sustainability commitment progress and updates. https://blogs.microsoft.com/blog/2021/07/14/made-to-measure-sustainability-commitment-progress-and-updates/, 2021.

[22] Justine Calma. Microsoft is changing the way it buys renewable energy. https://www.theverge.com/2021/7/14/22574431/microsoft-renewable-energy-purchases, 2021.

[23] Ross Koningstein. We now do more computing where there's cleaner energy. https://blog.google/outreach-initiatives/sustainability/carbon-aware-computing-location/, 2021.

[24] L. Lin and A. Chien. Characterizing stranded power in the ERCOT in years 2012-2019: A preliminary report. *University of Chicago CS Tech Report*, 2020.

[25] Sulav Malla, Qingyuan Deng, Zoh Ebrahimzadeh, Joe Gasperetti, Sajal Jain, Parimala Kondety, Thiara Ortiz, and Debra Vieira. Coordinated priority-aware charging of distributed batteries in oversubscribed data centers. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 839–851. IEEE, 2020.

[26] Meta. Responsible supply chain. https://sustainability.fb.com/responsible-supply-chain/, 2022.

[27] Microsoft. How microsoft measures datacenter water and energy use to improve azure cloud sustainability. https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-us 2022.

[28] Iyswarya Narayanan, Di Wang, Anand Sivasubramaniam, Hosam K Fathy, Sean James, et al. Evaluating energy storage for a multitude of uses in the datacenter. In *2017 IEEE International Symposium on Workload Characterization (IISWC)*, pages 12–21. IEEE, 2017.

[29] Lyswarya Narayanan, Di Wang, Abdullah-Al Mamun, Anand Sivasubramaniam, and Hosam K Fathy. Should we {Dual-Purpose} energy storage in datacenters for power backup and demand response? In *6th Workshop on Power-Aware Computing and Systems (HotPower 14)*, 2014.

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

[31] Maud Texier. Timely progress towards around-the-clock carbon-free energy. https://cloud.google.com/blog/topics/sustainability/t-eacs-help-drive-around-the-clock-carbon-free-energy 2022.

[32] The U.S. Energy Information Administration. Renewable energy explained. https://www.eia.gov/energyexplained/renewable-sources/, 2021.

[33] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E. Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: The next generation. In *Proceedings of the Fifteenth European Conference*

on Computer Systems, EuroSys '20. Association for Computing Machinery, 2020.

[34] Kasia Tokarska and Damon Matthews. Refining the remaining 1.5c 'carbon budget'. https://www.carbonbrief.org/guest-post-refining-the-remaining-1-5c-carbon-budget, 2021.

[35] U.S. Energy Information Administration. Annual energy outlook. https://www.eia.gov/outlooks/archive/aeo21/, 2021.

[36] Adam Wierman, Zhenhua Liu, Iris Liu, and Hamed Mohsenian-Rad. Opportunities and challenges for data center demand response. In *International Green Computing Conference*, pages 1–10. IEEE, 2014.