Data Center Demand Response for Sustainable Computing: Myth or Opportunity?

Ayse K. Coskun*, Fatih Acun*, Quentin Clark*, Can Hankendi*, Daniel C. Wilson†
*Electrical and Computer Engineering, Boston University, USA †Intel Corporation, Hillsboro OR, USA
{acoskun, acun, qtcc, hankendi}@bu.edu, daniel1.wilson@intel.com

***Abstract*—In our computing-driven era, the escalating power consumption of modern data centers, currently constituting approximately 3% of global energy use, is a burgeoning concern. With the anticipated surge in usage accompanying the widespread adoption of AI technologies, addressing this issue becomes imperative. This paper discusses a potential solution: integrating data centers into grid programs such as "demand response" (DR). This strategy not only optimizes power usage without requiring new fossil-fuel infrastructure but also facilitates more ambitious renewable deployment by adding demand flexibility to the grid. However, the unique scale, operational knobs and constraints, and future projections of data centers present distinct opportunities and urgent challenges for implementing DR. This paper delves into the myths and opportunities inherent in this perspective on improving data center sustainability. While obstacles including creating the requisite software infrastructure, establishing institutional trust, and addressing privacy concerns remain, the landscape is evolving to meet the challenges. Noteworthy achievements have emerged in the development of intelligent solutions that can be swiftly implemented in data centers to accelerate the adoption of DR. These multifaceted solutions encompass dynamic power capping, load scheduling, load forecasting, market bidding, and collaborative optimization. We offer insights into this promising step towards making sustainable computing a reality.**

***Index Terms*—data center, demand response, sustainability**

I. Introduction

Data centers are large partakers in global energy, and their demand is projected to grow with the escalating scale of computing services. Their non-regulated and enormous power usage raises sustainability issues, such as their carbon footprint and pressure on grid stability. This issue becomes even more concerning given the widespread adoption and increasing complexity of ML/AI workloads eagerly demanding more hardware and power. Given this immense and growing need, addressing the energy problems and the environmental impacts of large-scale computing systems becomes indispensable.

The grid is known to have tangible power supply fluctuations, accentuated by the increasing use of renewable energy. While wind power may counteract low solar power availability by increasing in the evenings, neither is completely reliable. Even in places with predictable renewable energy mixes, the net grid load often changes throughout the day resulting in a characteristic "duck curve" with a low net load mid-day and a high net load in the evenings. The significant power demand of data centers poses a further challenge to grid stability, due to both its mega-watt scale magnitude and dynamically changing characteristics. However, there are also emerging opportunities in the grid such as DR programs to address stability challenges in various ways (e.g., via monetary incentives) by requesting the demand side power consumers regulate their power consumption within predetermined timescales and rules. By engaging in such programs, consumers not only reduce their energy expenses but also contribute to grid stability and sustainability. Data centers are adaptable consumers uniquely well-positioned to participate in DR. They possess distinct capabilities to control power consumption rapidly and accurately by modulating utilization and the computing speed of their equipment.

Successful examples of participation in DR programs have been recently executed by data center operators [1]. However, there are major challenges that emerge while data centers operate under limited and varying power availability. This paper's goal is to highlight the promising recent developments as well as the key challenges. We present our takeaways for the current and future data center DR as follows:

- There is an evident but insufficiently investigated opportunity through DR for a mutualistic relationship between data centers and power grids to overcome supply challenges of renewable energy and the demands of exascale computing.
- Data centers are uniquely positioned contributors to DR programs due to the flexibility to finely and broadly adjust their power consumption.
- To boost the adoption of data center DR, it is essential to tackle facility concerns, particularly regarding the depreciation of system amortized value (i.e., by running at less than full load). While quality of service (QoS)-aware solutions are promising to overcome this hurdle, there is a need for their broader implementation, ensuring applicability across data centers of varying scales and diverse workload types.

II. Data Center Demand Response: Overview and Success Stories

There are a variety of DR programs offered by Independent System Operators (ISOs)[1] tailored to various objectives. Some programs aim to reduce the daily peak-to-average ratio of power consumption in the grid by encouraging the demand side to schedule their consumption to off-peak hours. On the other hand, in DR programs such as ancillary services, ISOs broadcast a fine granularity signal (e.g., changing every few seconds) that is translated by the participants to a dynamic power target to follow. Given the diverse range of DR programs,

[1]In this paper, ISO broadly denotes power providers and regulators, acknowledging alternate terms like Transmission System Operators in Europe.

979-8-3503-4859-0/DATE24/© 2024 EDAA

prior research affirms that data centers can actively engage in and benefit from those programs [2]–[4].

The industry's recent focus on DR is a promising sign for future enhancements in data center DR. Google announced pilot programs for DR in Europe, Asia, and the U.S., where they shift non-urgent compute tasks to reduce their power demand during peak hours in the grid [1]. Lawrence Livermore National Laboratory agreed to prepare up to 8 MW of demand flexibility for their electricity provider during a heat wave in Summer 2023 [5]. Despite the real-world examples illustrating data center DR being very recent, various research studies extending over a decade highlight the successful application of DR for data centers. The regulation service reserves program, a prominent DR program suitable for data centers due to potential monetary benefits (i.e., the value of provided flexibility being similar to the cost of consumed power), has been explored by developing models of data centers and their interaction with the ISO's regulation signal [2], [3].

III. Myths and Major Challenges

Despite the significant potential DR holds for increasing data center sustainability and promising examples already deployed, new solutions to address the challenges are still needed to promote widespread adoption of data center DR. The hesitation to implement DR stems from a combination of technical and social/political challenges in getting both data center and power provider operators to buy into DR programs. These challenges include both unwillingness from power providers to create and manage complex pricing arrangements, and concern from data centers that DR participation creates unacceptable risk when power management errors incur large economic penalties [4]. Another hurdle is the perception that DR would "waste" expensive hardware and compromise user experience [6].

A major challenge for further research into DR for data centers will be to debunk these myths by showing that DR can be employed such that it benefits both power suppliers and data centers, without compromising the QoS. Overcoming this latent resistance will require methods that are (1) robust enough to meet QoS constraints, (2) flexible to the needs of data centers of different scales and operational goals, and (3) based on accurate models that represent a real data center operation.

There is ongoing research into meeting each of these requirements. An overview of data center DR approaches is shown in Figure 1. One line of work treats data center control as a queuing problem, which enables designing efficient DR simulators for data centers [2]. Such a simulator, which has flexible parameters to approximate data centers of any scale or objective, can then be used to optimize DR parameters (e.g., server power, reserve capacity, etc.) to minimize an objective incorporating several cost factors as well as QoS constraints.

IV. Outlook

Enlightened by the recent real-world deployments of data center DR, the expectation for it to become widespread in the near future is elevating. Data centers present a promising opportunity for active participation in DR with their unique ability to precisely and rapidly control power consumption by

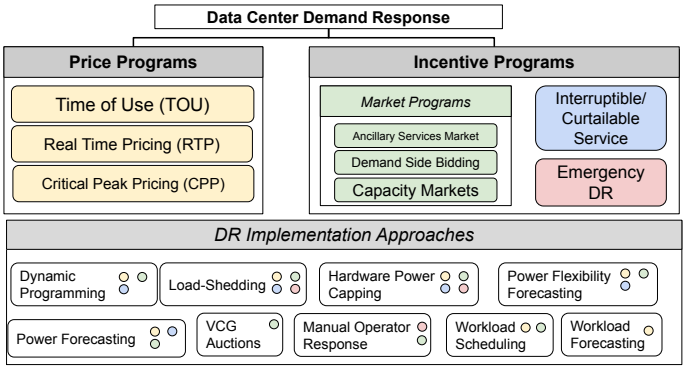

Fig. 1. An overview of the different forms of DR approaches for data centers that have been investigated in the literature.

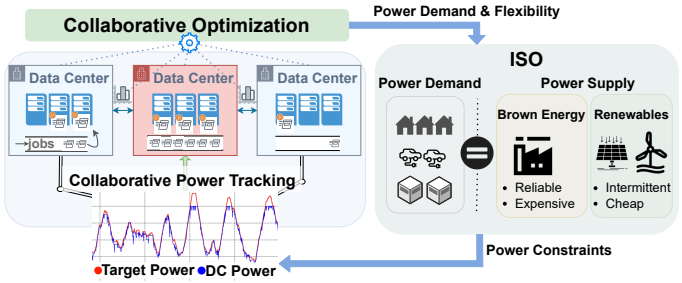

Fig. 2. Collaborative optimization of multiple data centers for DR allows providing an aggregated power demand and flexibility to the grid. Having a common power budget and a collaborative power-tracking mechanism results in more resilient system performance.

applying hardware power capping, workload scheduling, and resource allocation methods.

Addressing the challenges tied to the risk of performance degradation under power constraints is still a priority to expedite for broader adoption of DR. One promising direction is designing ML-powered methods that can rapidly change DR program predictions (such as power consumed and provided reserves) given dynamically changing load and QoS targets. *Collaboration* of multiple data centers poses another novel opportunity to absorb the risks during DR participation. Not only increasing the sustainability impact as a bigger entity, but collaboration will also help to exploit complementary characteristics (e.g., load, capacity, utilization) of data centers to tackle performance and power-related challenges (Figure 2).

References

[1] V. Mehra and R. Hasegawa, "Using demand response to reduce data center power consumption — google cloud blog," Oct 2023.

[2] Y. Zhang, D. Wilson, I. C. Paschalidis, and A. K. Coskun, "HPC data center participation in demand response: An adaptive policy with QoS assurance," *IEEE Trans. on Sustainable Comp.*, vol. 7, no. 1, pp. 157–171, 2022.

[3] W. Wang, A. Abdolrashidi, N. Yu, and D. Wong, "Frequency regulation service provision in data center with computational flexibility," *Applied Energy*, vol. 251, p. 113304, 2019.

[4] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response," in *IEEE International Green Computing Conference*, pp. 1–10, 2014.

[5] J. Kwan, "Climate change threatens supercomputers," *Science (New York, NY)*, vol. 378, no. 6616, pp. 124–124, 2022.

[6] A. Clausen, G. Koenig, S. Klingert, G. Ghatikar, P. M. Schwartz, and N. Bates, "An analysis of contracts and relationships between supercomputing centers and electricity service providers," in *Workshop Proceedings of the Intl. Conference on Parallel Processing*, ICPP Workshops, 2019.