Conductor: A Collaboration Framework for Multi-Data-Center Demand Response

Fatih Acun, Ioannis Ch. Paschalidis, Ayse K. Coskun
Boston University
Boston, USA
acun@bu.edu, yannisp@bu.edu, acoskun@bu.edu

2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC) | 979-8-3315-0786-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGSC64514.2024.00013

***Abstract*—Power consumption of data centers is rapidly becoming more prominent as the demand for computation increases. Next-generation systems are expected to require significantly more power, making it essential to design them to operate under power constraints to achieve sustainability goals. Power utilities are responsible for constantly providing reliable power and this task becomes harder as a higher amount of renewables are integrated into the grid. Demand response (DR) programs are promising solutions to maintain grid reliability by exploiting the flexibility of power consumers, such as data centers.**

While prior research explores data center participation in DR, real-world examples are still limited due to the risks of operating under power constraints, such as quality-of-service (QoS) violations. To provide greater flexibility in power consumption while improving data centers' ability to meet QoS targets, we propose *Conductor*, a novel framework that coordinates the participation of multiple data centers in DR, increasing their resilience to operate under power constraints without requiring any inter-data-center workload migration. Conductor assigns dynamic power targets to data centers based on their real-time QoS information and mitigates the risks of joining DR programs by recovering the QoS violations of jobs while achieving up to 78% better tracking of the power targets compared to individual data center DR participation.

***Index Terms*—data center collaboration, demand response, data center power management**

I. Introduction

The data center industry has been growing dramatically since the demand for computation is elevating with the recent advancements in AI and industrial workloads. The most powerful systems today consume substantial amounts of power. For instance, Frontier, the top-ranked computer in the TOP500 list [1], has a peak power consumption of 22.7 MW. By 2030, data centers are expected to account for 8% of total power consumption in the US, up from 3% in 2022 [2]. The increased size and number of data centers pose important sustainability challenges due to their huge power and energy consumption. This growth indicates the urgent need for action by data centers to transform their operations to become more sustainable.

Power utilities, in tandem, are restructuring their power supply sources with ambitious goals to decrease their carbon footprint by heavily relying on renewables. CAISO, as the most renewable-heavy utility in the US, plans to achieve a carbon-free power system by 2045 [3]. The intermittent nature of renewable energy amplifies the existing challenge of balancing the supply and demand in the grid. As a solution, independent service operators (ISOs) offer Demand Response (DR) programs to coordinate with the flexible power consumers in the grid to help maintain the balance. DR refers to the demand side regulating its power depending on the specific program requirements (e.g., [4], [5]). As data centers are large-scale and relatively flexible power consumers, they possess the potential to have a high impact on helping the ISOs by joining DR programs. By applying grid-aware workload scheduling and power management methods, they can quickly adjust to the power targets determined by DR programs.

Data center DR participation has been explored over a decade pointing out the monetary benefits that data centers can have by reducing their power costs [6]. Although recent real-world examples of data center DR are promising [7], [8], practical applications are still limited due to several reasons: (1) the risks of user job quality-of-service (QoS) violations while operating under power constraints, (2) the lack of ability to track the enforced power target. This paper argues that collaboration of multiple data centers for DR is a promising solution that can reduce the risks of QoS violations and enable lower energy costs by better tracking the power targets using the joint flexibility of multiple data centers at runtime (e.g., in load, power consumption, and utilization). Collaboration enhances grid flexibility, enabling sustainable growth and more renewable integration by uniting data centers into a cohesive entity with improved power tracking capability.

This paper introduces *Conductor*, a collaborative framework designed to enable multiple data centers under the same ISO to collectively provide flexibility to the power grid through participating in DR programs. Conductor allocates power targets to collaborating data centers by monitoring their real-time QoS metrics. Through collaboration, data centers can reduce the risks of QoS violations of their workloads and provide better tracking of the dynamic power targets compared to scenarios in which data centers individually join DR programs. **Key contributions** of this paper are as follows:

- A QoS-aware multi-data-center power balancing policy that improves system performance by mitigating QoS violations for all the collaborating data centers without implementing any inter-data-center workload migration.
- Improved tracking capability of the power target introduced by the ISO.
- An experimental evaluation of collaboration scenarios

979-8-3315-0786-2/24/$31.00 ©2024 IEEE
DOI 10.1109/IGSC64514.2024.00013

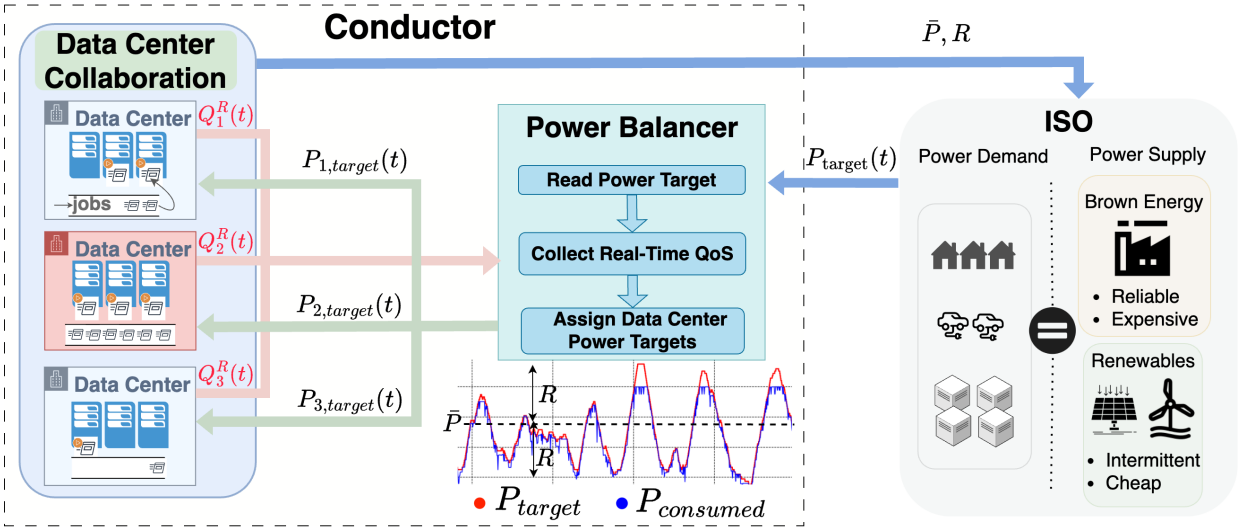


Fig. 1: Architecture diagram for Conductor. Data center collaboration provides its total average power consumption forecast $\bar{P}$ and reserve R bids to the ISO. Through the demand response execution over 1 hour, the power balancer distributes the collaborative power target, $P_{i,target}(t)$, to each data center based on the collected real-time QoS, $Q_i^R(t)$, at each time step.

with various data center configurations and workload properties through data center DR simulations.

Across a variety of scenarios with different numbers of data centers, utilization levels, and a wide range of workloads, we show that Conductor mitigates all the QoS service violations of each data center and simultaneously provides up to 78% better tracking of the collaborative target power compared to individual data center DR participations. To the best of our knowledge, our framework is the first to introduce power cap balancing among data centers through collaboration to track a joint total power target determined by a DR program.

The rest of the paper starts with a discussion of related work in Section II, followed by a description of our framework in Section III. We present our experimental methodology in Section IV. We provide our results in Section V, followed by a discussion in Section VI. We conclude in Section VII.

II. Related Work

Data center sustainability remains a critical issue as their power consumption grows significantly with the latest power-demanding applications and the deployment of newer systems. Addressing the carbon footprint of data centers has been a major concern. Various studies explore the flexibility of data centers on workload scheduling to execute the workloads and align the power consumption with the time-varying carbon intensity [9], [10]. Another crucial factor in data center sustainability is their impact on power grids since they influence grid stability as large-scale consumers. Their flexibility on power consumption allows them to cooperate with ISOs through DR programs. Data center DR participation methods explore workload scheduling and power management techniques to adjust the power consumption of data centers in response to grid objectives [6], [11], [12]. While participating in DR, a substantial priority for data centers is to satisfy the QoS requirements of their workloads, and QoS-aware methods are developed to address this need [13], [14].

The collaboration of multiple data centers has been studied for different goals such as carbon footprint and DR participation. To minimize the operational carbon footprint of data centers over multiple regions, Yang et al. [15] propose a spatio-temporal workload migration method to adjust data center power to align spatially with the renewable energy available across different geographical regions, and temporally with the time-varying local renewable generation. Lin et al [16] implement a load-balancing mechanism to minimize the energy cost and carbon footprint of geographically dispersed data centers. Similarly, studies explore workload migration to avoid energy curtailment [17] and minimize locational marginal carbon emissions [18]. DR participation is another area in which collaborative approaches are utilized. Lin and Guo [16] use a coalitional game theory method for data centers to collaboratively join capacity bidding DR programs and mitigate the unreliability of data centers' DR capability due to their random workload arrivals. Moghaddam et al. [19] propose a cloud federation approach using spatial workload migration and find the optimal workload allocation of cloud providers over different locations of the world. Another approach for multi-data center demand response utilizes the co-optimization of grid aggregators and multiple data centers that cooperatively operate using workload migration by providing QoS guarantees [20].

While some of the related studies enable collaboration by workload migration, it is not always a feasible option due to its time and energy costs as well as the challenges of executing workloads on different platforms. Our approach is unique among the related work since it does not require workload

migration and mitigates the QoS violations by providing flexibility for data centers over the collectively subscribed power target. In other words, instead of migration of workloads, we allow the data centers that need more computation capacity to consume more power dynamically.

III. A Collaborative Data Center Framework for Demand Response

Our aim for designing Conductor is to allow multiple data centers to participate in demand response programs collaboratively to reduce the risk of QoS violations and closely track the power targets of the ISO. Instead of each a single data center joining DR programs independently, we present a collaborative approach in which a data center collaboration interacts with the ISO through Conductor's QoS-aware power balancer. Conductor receives the power target, $P_{target}(t)$, from the ISO and distributes it to data centers in the collaboration based on the monitored real-time QoS metric, $Q^R(t)$. Figure 1 sketches the architecture diagram for Conductor.

A. Background on Demand Response

Power utilities offer various DR programs tailored for a wide range of problems in the grid. One such program is Emergency DR [4], where ISOs request power cuts from large consumers during severe conditions in the power grid caused by extreme weather, wildfires, and similar events. Another type of DR program, dynamic pricing, incentivizes the demand side to avoid peak hours by providing a varying price of electricity over time.

Regulation Service Reserves (RSR) [5] is a DR program in which power utilities broadcast a regulation signal $y(t) \in [-1, 1]$ at a certain frequency (e.g., every 4 seconds) to guide the power consumption of the demand side participants. We use RSR for data center DR since data centers can rapidly adjust their power consumption in response to the frequent regulation signal. To join RSR for a future period, data centers need to provide their forecasts of average power consumption $\bar{P}$, and reserve capacity R, which defines the amount of flexibility above and below $\bar{P}$. Their power target at time t, $P_{target}(t)$, is then defined in Equation (1). In RSR, participants are incentivized by discounting their energy use by their offered reserve amount of R, such that a larger R results in a reduced monetary cost. Similarly, they are penalized for their average tracking error, $\bar{\epsilon}$. The monetary cost of electricity use for time interval T (e.g., 1 hour) is calculated as in Equation (2), where Π^P, Π^R, and Π^ϵ denote the constants for the monetary price coefficients. Equation (3) defines the tracking error at time t as the absolute difference of the power consumption $P(t)$ with $P_{target}(t)$ relative to R. In this work, we constrain the tracking error to be less than a threshold of 0.3 for 90% of the time.

$$P_{target}(t) = \bar{P} + y(t)R, \tag{1}$$

$$\left(\Pi^P \bar{P} - \Pi^R R + \Pi^\epsilon R\bar{\epsilon}\right) T, \tag{2}$$

$$\epsilon(t) = \frac{|P(t) - P_{target}(t)|}{R}. \tag{3}$$

B. Our Data Center Model

For intra-data-center workload scheduling and power management, we use AQA, an Adaptive Policy with QoS Assurance for data center demand response [14]. AQA's data center model poses multiple job queues for different job types and assigns weights w to each queue to determine how many servers to allocate for each job type and satisfy the QoS constraints. It controls the cluster-level power consumption by idling/activating the servers and applying power caps to active servers based on the power target. In this work, we execute AQA for internal scheduling and power-capping decisions for each data center. The AQA framework also provides a gradient descent optimization method, which is done prior to demand response execution starts, to estimate optimal demand response parameters $\bar{P}$, R, and job queue weight parameters w for a single data center. The objective function for AQA optimization includes the monetary cost and QoS degradation costs as:

$$\begin{aligned} C \;=\; & \left(\Pi^P \bar{P} - \Pi^R R + \Pi^\epsilon R\bar{\epsilon}\right) \times \mathrm{T} \\ & + \beta \sum_j \mathrm{SoftPlus}\left(\rho\left(\mathrm{Prob}[Q^j - Q^j_{th}] - \delta^j\right)\right), \end{aligned} \tag{4}$$

where Q^j is the QoS degradation, and Q^j_{th} is the relative QoS threshold for job type j. The term $\ln(1 + e^x)$ is used as the SoftPlus function to penalize QoS degradations, which are defined probabilistically with a limit of $\delta^j = 10\%$ allowing only a small portion of jobs to violate the QoS thresholds. β and ρ are weighing factors to balance the monetary cost and unitless QoS cost. The QoS degradation, Q^j, defined in Equation (5), is formulated as the difference between a job's total elapsed time from job submission to completion (T_{so}), and minimum execution time (T^j_{min}) without any power caps:

$$Q^j = \frac{T_{so} - T^j_{min}}{T^j_{min}}. \tag{5}$$

C. Multi-Data-Center Power Balancer

Power sharing within the DC collaboration is acquired by a central power balancer that distributes the available power to DCs by introducing individual power targets, decomposing the power enforced by the power utility fairly. The power balancer enforces the fairness criteria by considering the monitored real-time QoS metric, defined as a function of time t as follows:

$$Q^R(t) = \frac{1}{|j|} \sum_j \frac{q_{0.9}\left(\dfrac{T_{\text{wait}}(t) + T_{\text{exec}}(t)}{T^j_{\text{min}}}\right)}{Q^j_{\text{th}}}, \tag{6}$$

where T^j_{min} refers to the minimum job execution time without any power cap applied, and Q^j_{th} defines the QoS threshold for job type j. For timestep t, $T_{wait}(t)$ and $T_{exec}(t)$ denote the job wait time in the queue and execution time after being scheduled. The $Q^R(t)$ metric is aggregated over j different job types by calculating the average of 90^{th} percentiles ($q_{0.9}$) for each job type. During demand response execution, each data

center calculates $Q^R(t)$ at every second by the current state of their jobs' progress and delivers it to the power balancer.

The power balancer calculates the specific power targets for each data center in the collaboration. Equation (7) determines the power target at time t for data center i by distributing the total target power as follows:

$$P_{i,target}(t) = P_{target}(t)\left(\tau\frac{Q_i^R(t)}{\sum_{k=1}^{S} Q_k^R(t)} + \phi\frac{\bar{P}_i}{\sum_{k=1}^{S}\bar{P}_k}\right), \quad (7)$$

where $P_{target}(t)$ shows the total collaborative power target calculated as in Equation (1), and $P_{i,target}$ denotes power target assigned to data center i (DC_i). The quantities Q_i^R and $\bar{P}_i$ are the real-time QoS value and the average power bid for DC_i. $Q_i^R(t)$ and $\bar{P}_i$ are normalized by their corresponding total sums for S data centers in the collaboration. The parameters τ and ϕ are balancing factors and set to 0.5 to give equal importance to average power consumption and the real-time QoS values of data centers.

IV. Experimental Methodology

Our experiments assess Conductor's capability compared to the scenarios in which data centers individually participate in DR programs. We evaluate our framework based on two criteria: (1) QoS constraints of the jobs, (2) the tracking error of the power target introduced by the DR program. We use an in-house data center DR simulator to do experiments with large-scale data centers. The simulator input consists of the experiment configurations such as data center size, job types and their power performance properties, and data center utilization. The jobs used in this study, are NAS parallel benchmarks [21], and their power performance properties are collected by profiling them in [14] using Massachusetts Green High-Performance Computing Center [22]. The benchmark applications we use include `is`, `ep`, `cg`, `mg`, `ft` and we run the benchmarks with inputs `C` and `D` with different numbers of threads and using up to 8 nodes. In this paper, we format the job configurations as `<benchmark_name>.<input>.<num_threads>`.

To represent data centers with various configurations, we explore data centers with high/low utilization, tight/slack QoS constraints, and high/low power-consuming applications. Table I describes the workload traces that we use to simulate different data centers. More detail for the generation of those workloads can be found in [14]. We build different types of experiment scenarios with different numbers of data centers joining the collaboration having varying configurations. Table II summarizes the scenarios that cover the collaboration experiments in this work. Each data center in our experiments comprises 1,000 servers, representing a mid-scale facility. This size is sufficiently large to conceptually demonstrate the capabilities of our framework.

We first run gradient descent optimization of the objective function, C, for each data center to independently find the optimal values for $\bar{P}_i$ and R_i. After the data center collaboration delivers the total bids, $\bar{P}$ and R, to the ISO, we start the collaborative DR simulation with Conductor. To implement the multi-data-center power balancer, we execute our data center simulator for S data centers using S parallel processes. For each time step during the execution, we use a shared-file-based approach between all the simulator processes to implement the communication between the data centers and the power balancer module.

TABLE I: Workload traces and their properties used in the scenarios.

Workload Trace	Property	Description
W_{LP}	Low Power	Low power consuming jobs
W_{HP}	High Power	High power consuming jobs
W_{TQ}	Tight QoS	Jobs with $Q_{\text{thres}} \leq 5$
W_{SQ}	Slack QoS	Jobs with $Q_{\text{thres}} > 5$
W_{LU}	Low Util	Average system utilization is 25%
W_{HU}	High Util	Average system utilization is 90%

TABLE II: List of collaboration scenarios used in experiments.

Scenario	# of Data Centers	Workload Traces
S_1	2	$W_{(LP,HP)}$
S_2	2	$W_{(TQ,SQ)}$
S_3	2	$W_{(LU,HU)}$
S_4	4	$W_{(TQ,SQ,LU,HU)}$
S_5	6	$W_{(LP,HP,TQ,SQ,LU,HU)}$

V. Results

In this section, we present the results of collaborative experiments for data center DR participation with Conductor, compared to individual DR participation.

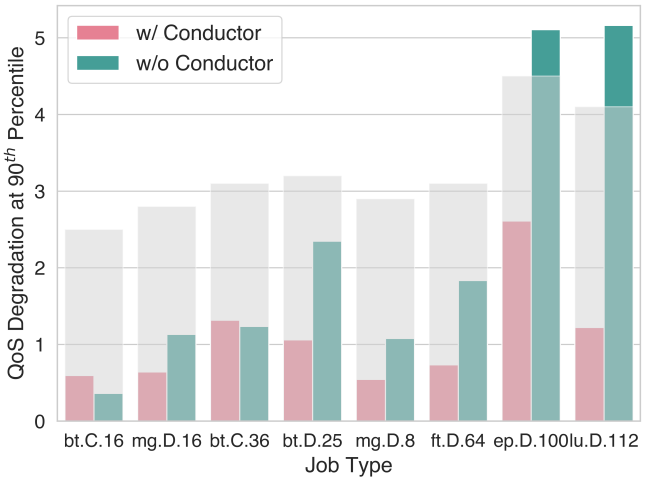
A. Avoiding QoS Violations with Collaboration

One important risk of operating under a power constraint for a data center is not satisfying the QoS requirements of their jobs. As the power budget goes down, the slowdown that each job experiences becomes greater, and so is the probability of QoS violations. Conductor assigns higher power targets for data centers with higher real-time QoS (Q^R) values and therefore prevents QoS violations. Figure 2 shows the QoS degradation results for scenario S_2, in which we execute 2 data centers with workload traces W_{TQ} and W_{SQ}. Experiments without Conductor show QoS violations for two job types `ep.D.100` and `lu.D.112`, in the tight-QoS data center, and one violation in the slack-QoS data center for `is.D.32`, exceeding their QoS thresholds as indicated by the gray-shaded bars. If the collaborative execution is activated with Conductor, all the QoS violations are avoided for both data centers by bringing the QoS degradations within thresholds. Summarized results for all scenarios are shown in Table IV.

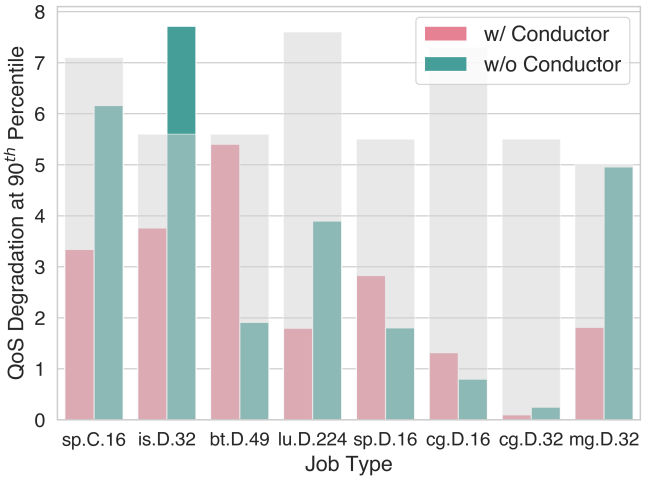
In addition to avoiding QoS violations, we also compare the average of Q^R values across all data centers. We calculate the mean of Q^R defined in Equation (6) for each scenario after the simulations are completed. The results in Table III show that the mean Q^R value is improved in every scenario.

B. Tracking the Power Target

In DR participation, data centers need to track the power targets strictly or they will face the risk of increased monetary penalties for tracking errors and even lose their DR contracts



(a) DC_1 running Tight QoS workload mix W_{TQ}.



(b) DC_2 running Slack QoS workload mix W_{SQ}.

Fig. 2: QoS degradation results at 90^{th} percentile for each job type for scenario S_2. Gray-shaded bars show the QoS thresholds for each job type. Violations exceeding the thresholds are recovered for both data centers with Conductor.

with the ISO. By participating in DR alone, data centers might not provide good tracking capabilities under certain conditions such as tight QoS constraints, too low/high utilizations, or unexpected workload spikes and performance variations. In such scenarios, Conductor exploits complementary behaviors of the data centers in collaboration to match the power target.

TABLE III: Monetary cost and mean Q^R results. Lower values of mean Q^R indicate less QoS degradations.

Scenario	w/o Conductor		w/ Conductor	
	Cost ($)	Mean Q^R	Cost ($)	Mean Q^R
S_1	45.29 $	0.33	45.26 $	0.27
S_2	45.09 $	0.60	45.07 $	0.37
S_3	44.02 $	0.40	43.60 $	0.26
S_4	89.12 $	0.50	88.53 $	0.33
S_5	134.40 $	0.44	133.83 $	0.33

TABLE IV: QoS violation recoveries with Conductor. Only the data centers with non-zero violations in 'w/o Conductor' are shown for simplicity.

Scenario	Data Center	% of Job Types Violating QoS	
		w/o Conductor	w/ Conductor
S_2	DC_1 (W_{TQ})	25.0%	0.0%
	DC_2 (W_{SQ})	12.5%	0.0%
S_3	DC_2 (W_{HU})	25.0%	0.0%
S_4	DC_1 (W_{TQ})	25.0%	0.0%
	DC_2 (W_{SQ})	12.5%	0.0%
	DC_4 (W_{HU})	25.0%	0.0%
S_5	DC_3 (W_{TQ})	25.0%	0.0%
	DC_4 (W_{SQ})	12.5%	0.0%
	DC_6 (W_{HU})	25.0%	0.0%

We present the power tracking ability of the collaboration for scenario S_3 in Figure 3. For illustration purposes, we trigger Conductor at the end of the first hour. As seen in the top of Figure 3, the data center collaboration can closely track the collaborative power target as each data center tracks their power targets assigned by the power balancer shown in the second subfigure. Power targets are assigned based on $Q^R(t)$ metric shown in the bottom subfigure. We include the tracking errors at the 90^{th} percentile in Figure 4 with and without Conductor for the same scenario S_3. Without Conductor, the data center with W_{LU} violates the tracking error constraint of 0.3, reaching a value of 0.54. In contrast, with Conductor, the collaborative tracking error stays within the constraint at 0.27 as shown with the blue bar. Also, Table III shows that there is a minor improvement in the monetary cost up to %0.96 across all scenarios. The reason for this improvement is attributed to achieving less penalty for lower tracking errors. Improving tracking error has a minor effect on the monetary cost since it holds smaller values compared to $\bar{P}$ and R, which control the other terms of the monetary cost defined in Equation (2).

C. Analysis of the Size of the Collaborating Data Center Cohort

In this set of experiments, we test Conductor for collaborations with more than 2 data centers by including the scenarios S_4, and S_5. We present tracking errors of Conductor compared to individual DR participations of different scenarios in Figure 4. The results certify that Conductor can still achieve good tracking results by providing a collaborative tracking error within constraints. We also observe that tracking capability improves with the increasing number of data centers in the collaboration.

Table IV presents the results for the scenarios in which QoS violations are observed for individual DR participations. We provide the percentage of job types that violate their QoS threshold in a data center and violations in all scenarios are recovered with Conductor, proving its ability to operate with collaborations having a higher number of data centers.

VI. Discussion

In this work, we design Conductor for data center collaboration to participate in DR programs by providing better tracking capability of the power targets and simultaneously

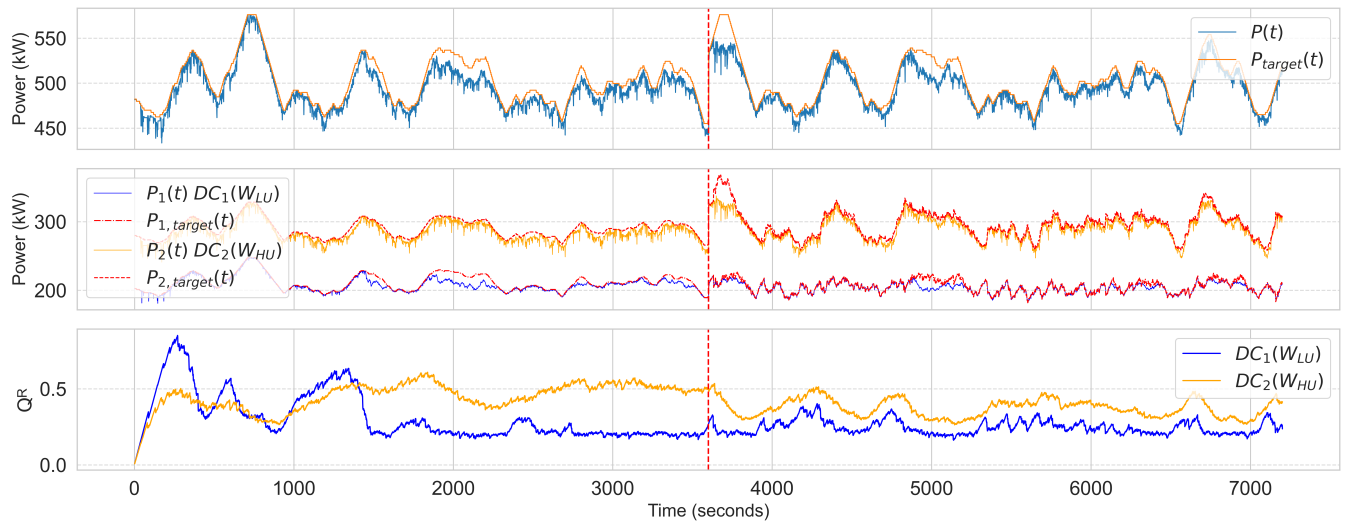


Fig. 3: Power tracking as a collaboration for scenario S_3. Three subplots show the collaborative power tracking, the tracking of power targets assigned by the power balancer, and the collected $Q^R(t)$ metric over time, respectively. Conductor is activated at the end of the first hour, as indicated by the vertical dashed line.

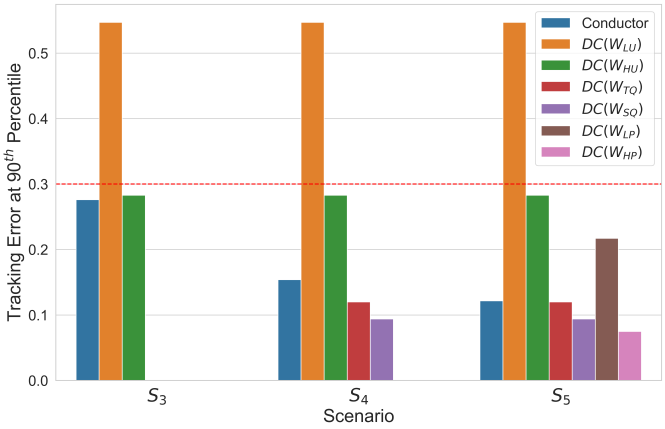


Fig. 4: Comparison of tracking errors at 90^{th} percentile for individual DR participation of data centers to collaboration with scenarios of varying numbers of data centers. The horizontal dashed line shows the tracking error constraint at 0.3.

recovering from QoS violations. Our framework targets the data centers joining demand response programs in the same ISO region. Since those data centers are located in the same region, they can collaborate by only properly arranging their power consumption, without a need for workload migration. While we achieve improvements in monetary cost, it is important to underline that collaborative data center demand response methods hold promising potential for the impact of data centers on grid stability. Since data centers are already large-scale power consumers, enabling them to operate in collaboration will escalate their impact when responding to the ISO's requests in DR programs.

Our framework stands as a research prototype that demonstrates a way to implement such collaborative methods by sharing minimal information between data centers. Our results show the opportunity to use the joint flexibility of multiple data centers stemming from their workloads with different characteristics during DR participation. The real-world adoption of such collaborative methods is relatively easy to implement for data centers owned by the same company. However, collaboration among data centers owned by different companies may bring additional challenges due to privacy concerns for sharing information such as QoS, and power consumption. A potential solution to such challenges is using privacy-preserving methods (e.g., differential privacy) to prevent data centers from inferring other data centers' confidential information.

VII. Conclusion and Future Work

In this study, we provide Conductor, a multi-data-center collaboration framework that provides flexible capacity to the power grid through DR participation. By sharing minimal information between data centers in the collaboration, Conductor utilizes a QoS-aware power balancer to reduce the risk of QoS violations and achieves lower power costs by better tracking the power target by the ISO.

As future work, our goal is to explore collaborative optimization methods to significantly decrease monetary costs of energy by offering lower average power and higher reserve to the ISO using the increased flexibility provided by the Conductor. To address the potential privacy concerns of different companies, we aim to extend our approach to information sharing with a privacy-preserving method.

REFERENCES

[1] TOP500.org, "Top500: The list." https://www.top500.org/lists/top500/, 2024. Accessed: 2024-06-10.

[2] Goldman Sachs, "Ai poised to drive 160% increase in power demand." https://www.goldmansachs.com/intelligence/pages/AI-poised-to-drive-160-increase-in-power-demand.html, 2024. Accessed: 2024-06-10.

[3] California Independent System Operator (CAISO), "2022-2026 strategic plan." https://www.caiso.com/Documents/2022-2026-Strategic-Plan.pdf, 2022. Accessed: 2024-06-10.

[4] California Public Utilities Commission, "Emergency Load Reduction Program." https://www.cpuc.ca.gov/industries-and-topics/electrical-energy/electric-costs/demand-response-dr/emergency-load-reduction-program. Accessed: 2024-06-29.

[5] PJM Interconnection, "Manual 12: Balancing operations." https://pjm.com/~/media/documents/manuals/m12.ashx. Accessed: 2024-06-10.

[6] 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.

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

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

[9] A. Radovanović, 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," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1270–1280, 2023.

[10] F. Yang and A. A. Chien, "Zccloud: Exploring wasted green power for high-performance computing," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 1051–1060, 2016.

[11] 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.

[12] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: avoiding the coincident peak via workload shifting and local generation," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, p. 341–342, jun 2013.

[13] A. Jahanshahi, N. Yu, and D. Wong, "Powermorph: Qos-aware server power reshaping for data center regulation service," *ACM Trans. Archit. Code Optim.*, vol. 19, aug 2022.

[14] 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, pp. 157–171, 2022.

[15] T. Yang, H. Jiang, Y. Hou, and Y. Geng, "Carbon management of multi-datacenter based on spatio-temporal task migration," *IEEE Transactions on Cloud Computing*, vol. 11, no. 1, pp. 1078–1090, 2023.

[16] W.-T. Lin, G. Chen, and H. Li, "Carbon-aware load balance control of data centers with renewable generations," *IEEE Transactions on Cloud Computing*, vol. 11, no. 2, pp. 1111–1121, 2023.

[17] J. Zheng, A. A. Chien, and S. Suh, "Mitigating curtailment and carbon emissions through load migration between data centers," *Joule*, vol. 4, no. 10, pp. 2208–2222, 2020.

[18] J. Lindberg, L. Roald, and B. Lesieutre, "The environmental potential of hyper-scale data centers: Using locational marginal co2 emissions to guide geographical load shifting," in *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 2021.

[19] M. M. Moghaddam, M. H. Manshaei, W. Saad, and M. Goudarzi, "On data center demand response: A cloud federation approach," *IEEE Access*, vol. 7, pp. 101829–101843, 2019.

[20] Y. Zhang, A. Tsiligkaridis, I. C. Paschalidis, and A. K. Coskun, "Data center and load aggregator coordination towards electricity demand response," *Sustainable Computing: Informatics and Systems*, vol. 42, p. 100957, 2024.

[21] NASA Advanced Supercomputing Division, "NASA Parallel Benchmarks." https://www.nas.nasa.gov/software/npb.html. Accessed: 2024-06-29.

[22] Massachusetts Green High Performance Computing Center, "MGHPCC Website." https://www.mghpcc.org/. Accessed: 2024-06-29.