Site-Wide HPC Data Center Demand Response

Daniel C. Wilson
Boston University
Boston, USA
danielcw@bu.edu

Ioannis Ch. Paschalidis
Boston University
Boston, USA
yannisp@bu.edu

Ayse K. Coskun
Boston University
Boston, USA
acoskun@bu.edu

***Abstract*—As many electricity markets are trending towards greater renewable energy generation, there will be an increased need for electrical grids to cooperatively balance electricity supply and demand. Data centers are one large consumer of electricity on a global scale, and they are well-suited to act as a grid load stabilizer via performing "demand response."**

Prior investigations in this space have demonstrated how data centers can continue to meet their users' quality of service (QoS) needs by modeling relationships between cluster job queues, server power properties, and application performance. While server power is a major factor in data center power consumption, other components such as cooling systems contribute a non-negligible amount of electricity demand.

This work proposes using a simple site-wide (i.e., including all components of the data center) power model on top of QoS-aware demand response solutions to achieve the QoS benefits of those solutions while improving the cost-saving opportunities in demand response. We demonstrate 1.3x cost savings compared to QoS-aware demand response policies that do not utilize site-wide power models, and show similar savings in cases of severely under-predicted site-wide power consumption if 1.5x relaxed QoS constraints are allowed.

***Index Terms*—HPC, Demand Response, Power Usage Efficiency, Quality of Service**

I. Introduction

Electricity markets are trending toward supply mixes with increasing proportions of renewable energy sources. Most states in the U.S. have adopted renewable portfolio standards to increase their renewable energy generation, and 12 states have put plans into action to achieve 100% clean energy within the next 30 years [1]. The U.S. Energy Information Administration forecasts that 70% of the renewable energy supply will consist of solar and wind power in that time frame [2]. Solar and wind power supplies have time-varying availability, so there is a growing need for power management solutions that adaptively balance supply and demand.

Regulation service programs are one solution to manage imbalances between electricity supply and demand. In such programs, an electricity consumer offers regulation capacity to an independent service operator (ISO) in the grid. The offered service is a promise that the consumer will modulate power consumption to help the ISO balance electricity supply and demand. The consumer receives cost incentives depending on constraints specified by the ISO. For example, the PJM ISO has a program that pays for regulation service based on changes in hourly rates, on the magnitude of reserve offered by a consumer, and on the consumer's quick and accurate response to the ISO's requests for power consumption modulation [3]. In this work, we refer to a *bid* as the amount of average purchased power for an upcoming hour, and the reserve capacity offered for that hour.

Data centers are well-suited to offer regulation service because they consume a lot of electricity and they can quickly modulate their power consumption through job scheduling and server-level power management. An ACM tech brief estimates that data centers consumed 3% of the total energy supply in 2021, and are likely to demand more as power-hungry cryptocurrency and artificial intelligence workloads increase in popularity [4]. Prior works have demonstrated that data centers can effectively manage trade-offs between power consumption and application performance to meet system-wide power objectives [5], [6].

Previous work toward using data centers as reserve capacity in a smart grid has focused on enhancing trade-offs between quality-of-service (QoS) and server power consumption, such as by co-optimization of battery-based and server-workload-based power reserves [6], offering probabilistic guarantees on QoS degradation [7], and enhanced bidding strategies to exploit properties of long-running workload mixes observed in real-world data centers [8].

There is currently a gap in efforts to design QoS-aware demand response policies that operate at the site level (i.e., including all components of the data center). Beyond server power in a data center, cooling and power delivery infrastructure contribute toward significant additional power demand. In a 2021 survey, large data centers used on average over a third of their power consumption for non-server needs [9]. While prior work proposes methods to co-optimize frequency control and cooling system parameters [10], there isn't a single solution that integrates QoS-aware decisions along the entire path of bidding, job scheduling, and site-scoped power management. This paper aims to close the research gap by evaluating the opportunity to use site-level power consumption in demand response power management policies.

The **key contributions** in this work are as follows:

- We demonstrate that site-wide demand response participation in a data center enables lower operating costs than server-only demand response participation while using an existing QoS-aware demand response policy. We achieve 1.3x cost savings with similar QoS degradation

Research was partially funded by the Hariri Institute and the Institute for Global Sustainability at Boston University.

by making demand response policies aware of site-wide power consumption.

- We analyze the sensitivity of demand response cost savings to site-wide power model selection. In scenarios where there is low confidence in the site-wide power model, demand response policies can still reduce electricity costs, but may need to relax their QoS constraints.
- We show that batch job resource managers meet site-level demand response objectives and job-level QoS objectives by using a simple site power model on top of QoS-aware power managers that operate with job and server scope.

The remaining sections of this paper discuss related works for data centers in demand response, describe our simulation infrastructure, explain our experimental methodology, and evaluate our simulation results.

II. Related Work

There is broad interest in reducing the carbon impact and cost impact of data centers while maintaining quality of service. A survey of energy-aware Top500 supercomputing sites indicates that facilities are interested in job scheduling, resource management, and facility design strategies to meet their high-level power objectives [11]. Another survey reveals that a lack of QoS guarantees is a strong reason that some data centers do not consider demand response in their electricity procurement strategies [12].

Existing works about demand response in data centers often investigate QoS awareness without pursuing site-level power awareness. The Adaptive policy with Quality Assurance (AQA) [7] operates on server-level power metrics and controls to enable QoS-aware bidding and power management in demand response in HPC environments. The ECOGreen policy [6] focuses on QoS awareness in virtualized computing environments, and includes on-site power storage as a control.

Other work [10] explores site-level monitoring and control to meet target power levels in demand response programs. While that work does include QoS estimates as part of the bidding strategy, it does not include end-to-end QoS feedback from bidding through scheduling and power management.

Some efforts work toward sharing work across data centers to help balance electricity supply. For example, the Zero-Carbon Cloud project [13] relocates virtual computing workloads so that computing power demand can follow changing supply. That type of solution offers high capacity to match changing power supply, particularly when many small data centers are distributed near different energy supplies. However, that type of approach is not suitable for workloads that are expected to operate within a single HPC data center.

Our work augments cluster-level, QoS-aware demand response policies with a simple site-level power model to enable site-wide demand response participation of HPC sites.

III. Site-Wide Demand Response

Our high level goal is to identify the cost-saving opportunities of demand response bidding and power management policies in an HPC data center, while meeting QoS constraints of jobs in a work queue. This section describes the data center simulator and the demand response policies we use to analyze our cost-saving opportunities.

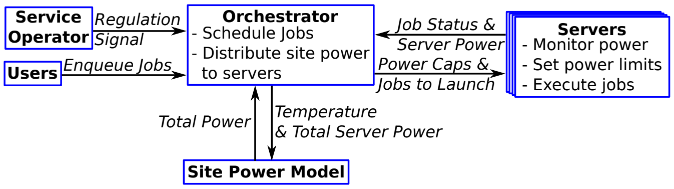

Fig. 1. The simulator places control logic in an orchestrator that schedules work and applies power limits to servers to satisfy user and independent service operator objectives.

A. Data Center Simulator

To evaluate the quality of demand response policies, we simulate a data center's work queues, power consumption across the data center, performance of running applications, and incoming regulation service control signals. As shown in Fig. 1, simulation logic resides in a central orchestrator, which acts as the data center's job scheduler and resource manager.

The overall execution flow of the simulator begins by loading application and data center properties, then executing the bidding policy, and lastly performing fixed-time-interval simulations of servers, the regulation signal, and the demand response policies for scheduling and power management.

For each one-second time step simulating data center servers, the simulator updates its queue of incoming job requests from users and reads the regulation signal from the independent service operator. The simulator updates the cluster-level view of server power consumption prior to executing scheduling policies and power-capping policies, which update control signals on the simulated servers.

Job submissions are randomly generated for each simulation, as a Poisson process for each job type. The average arrival rate of each job type is selected to evenly distribute a target data center utilization across all job types in a workload mix.

B. Demand Response Policy

We simulate the *AQA* policy [7] for demand response control decisions. In this work, we update the policy to make it aware of site-wide power consumption. We use a site-wide power model to inform the policy how much server power needs to increase or decrease to meet the site-wide power target, and the policy makes its QoS and server-power-aware control decisions in the same manner as described in its original work. This design choice ensures that existing policies can make site-level power control decisions by using a simple model of site-wide power consumption. We refer to the updated site-wide AQA policy as *AQA-SW* in this paper.

The AQA algorithm applies weights to individual job types in a workload mix. Those weights are used to influence which queued jobs get sent to servers, as well as how the power budget is distributed across active servers. Weights and bids are learned prior to evaluating each workload mix

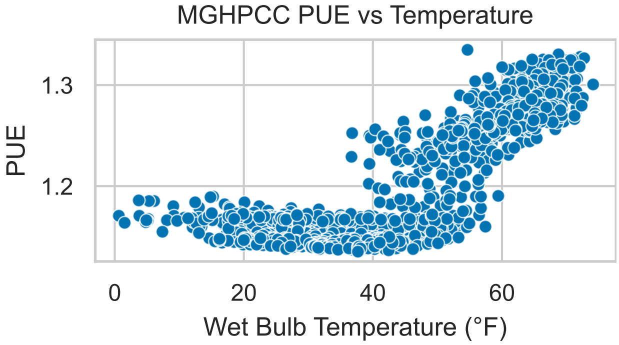

Fig. 2. PUE of the MGHPCC data center as a function of outdoor wet bulb temperature.

by running many simulations of a workload mix through a gradient descent search [7].

In this work, we use the same search used by AQA, but we run the AQA-SW policy in each gradient descent step. The search incorporates electricity cost, QoS constraints, and power-tracking constraints in the gradient descent cost function. Some search steps occasionally enter regions of violated constraints, then adjust in later steps to find viable bids and weights. We stop searching after 120 iterations of gradient descent, prune non-viable search steps, and select the bid and weights with the lowest cost.

C. Site Power Model

The simulator estimates power consumption at the site level by using a power usage efficiency (PUE) model. PUE is defined as PUE $= \frac{\text{Total Power}}{\text{IT Power}}$, so our simulator estimates total power as the product of the modeled PUE and the sum of power consumption across servers.

For this work, we construct a piecewise-linear model of PUE as a function of outdoor wet-bulb temperature, where hotter and more humid days will place a greater cooling load on the data center for a given computational workload. The PUE model is used with the definition of PUE $= \frac{\text{Total Power}}{\text{IT Power}}$ to translate between IT, non-IT, and total power where needed by the simulator. The model of PUE as a function of wet bulb temperature is:

$$\text{PUE}(x) = \begin{cases} A(x-B)+C & x < B \\ D(x-B)+C & x \geq B \end{cases}$$

The model's parameters are fit with least squares regression against power and wet bulb temperature data provided by the Massachusetts Green High Performance Computing Center (MGHPCC) [14]. The MGHPCC is a megawatt-scale computing center with hundreds of thousands of CPU cores and millions of GPU cores. We use data from 2018 to 2020, shown in Fig. 2. To work with the available data, we approximate PUE as the ratio of the data center's power consumption to the power consumed by all computing racks in the center clusters. The resulting model parameters are $A = 5\times10^{-12}$, $B = 42.1$, $C = 1.16$, and $D = 5.5 \times 10^{-3}$.

The model predicts the source data's PUE with 1.1% mean absolute error. We do not focus on developing a robust model in this work. Instead, we evaluate how server-only demand response policies meet their QoS objectives when they are augmented with a simple model to enable site-wide power awareness. We evaluate sensitivity to modeling error in Section V.

Power consumption of simulated servers follows a simple model based on activity. If no workload is executing on the server, the server is assumed to consume its idle power. Idle power is assumed to be 169 watts in our experiments as measured in previous experiments on a subset of servers in that cluster [7]. If a workload is executing on the server, then it is assumed to consume the lesser of the server's current power limit (as set by the orchestrator) and the current workload's maximum power consumption.

IV. Methodology

Our experiment plan consists of two stages. First, we evaluate the cost savings of a site-wide demand response policy in comparison to the savings from a server-power-only policy. Second, we evaluate the sensitivity of those cost savings with respect to the choice of site-wide power consumption model.

Both the cost-savings evaluation and the sensitivity analysis are performed over a range of simulated external wet bulb temperatures from $40°$F to $70°$F, in order to exercise data center efficiency levels between the corner cases of observations from the PUE model described in Section III-A.

The properties of our simulated data center are based on sampled data from real servers in the MGHPCC. As described in Section III-C, our site-wide power model is fit against site power data. The execution times and power trends of our simulated applications are sampled from application runs on Boston University's Shared Computing Cluster, which is hosted in the MGHPCC.

We select several workloads from the workload mixes in the work that introduced the AQA policy [7]. The selected workloads and their compositions of applications are shown in Table I. We pick workloads that exhibit mixed properties (W1), as well as corner cases such as low power (W8), high power (W9), low utilization (W12), and high utilization (W13).

A. Cost Savings

Our goal in the cost-savings experiments is to evaluate the cost opportunities that are offered by using site-wide demand response instead of server-only demand response in an HPC data center. We simulate the AQA policy for server-only demand response and the AQA-SW policy for side-wide demand response (both described in Section III-B).

1) Cost Model: We model the total cost of electricity as the estimated cost of an energy purchase, minus the value of the regulation service provided by the data center. Specifically, we model the hourly cost as

$$\text{Cost} = \Pi_P \bar{P} - \Pi_R R + \Pi_\epsilon R\epsilon,$$

where $\Pi_P \bar{P}$ indicates the cost of energy purchased for the hour, $\Pi_R R$ represents the value of reserve capacity offered for the hour, and $\Pi_\epsilon R\epsilon$ is the cost penalty for ϵ percent error

TABLE I

MEMBERSHIP OF APPLICATIONS WITHIN WORKLOAD MIXES. EACH APPLICATION FOLLOWS A PATTERN OF `<NAME>.<INPUT CLASS>.<PROCESS COUNT>`. THE m_j COLUMN SHOWS THE NUMBER OF SERVERS USED PER JOB. T_{min} (T_{max}) IS THE MINIMUM (MAXIMUM) PROCESSING TIME IN SECONDS, AND p_{max} (p_{min}) IS THE CORRESPONDING POWER CONSUMPTION OF A SERVER IN WATTS. Q_{thres} IS THE THRESHOLD FOR EACH WORKLOAD MIX'S ACCEPTABLE LEVEL OF QOS DEGRADATION.

App	m_j	T_{min}	p_{max}	T_{max}	p_{min}	Q_{thres}	W1	W8	W9	W12	W13
bt.C.16	1	73	339	86	249	2.5		✓			
mg.D.16	1	84	380	105	266	2.8			✓		
sp.C.16	1	54	375	62	267	7.1			✓		
is.D.32	3	42	249	42	241	5.6	✓	✓		✓	✓
bt.C.36	2	38	343	46	249	3.1		✓			✓
bt.D.49	2	551	391	729	250	5.6	✓		✓	✓	✓
ep.D.64	3	54	353	70	237	3.9			✓		
sp.D.100	4	343	399	381	264	3.3			✓		
ep.D.28	1	124	383	175	238	5.9			✓		
cg.C.4	1	28	238	28	239	4.0		✓			
bt.D.25	1	1022	402	1370	254	3.2			✓		
mg.D.8	1	141	297	151	258	2.9	✓	✓		✓	✓
is.D.4	1	122	204	123	194	7.3	✓			✓	✓
cg.D.16	1	743	326	823	253	7.3		✓			
ep.D.56	2	64	383	90	238	2.0					
cg.D.128	6	231	336	242	246	4.0		✓			
cg.D.32	3	364	281	390	246	5.5		✓			
ep.D.100	4	36	366	49	238	4.5	✓			✓	✓
is.D.64	4	27	287	28	228	3.1	✓			✓	
lu.D.112	4	164	405	222	251	4.1	✓			✓	✓
mg.D.32	2	49	378	58	266	5.0	✓		✓	✓	✓

in tracking the regulation signal. While this cost model is in line with prior work on data center demand response [7], the key difference is that we calculate the cost using *site-wide* power consumption.

2) Sources of Cost Savings: The site-wide demand response policy accounts for all the site's power in the above cost model. The average purchased power of the AQA and AQA-SW policies is derived from the total power consumed by the site. Note that although the cost equations appear the same across AQA and AQA-SW cases, the actual cost differs because the magnitude of R increases in the site-wide policy, since it is derived from both server and non-server power in that case.

As a baseline, we estimate non-demand-response cost based on the total energy consumed, times the cost per kilowatt-hour of energy. As a result, the baseline cost does not directly depend on the $\bar{P}$ and R cost components describe above, but the baseline cost does depend on the scheduling and power management decisions made in the data center.

B. Sensitivity Analysis

We analyze the demand response policy's sensitivity to the choice of site-wide power model. The purpose of this analysis is to identify how well the policies respond when the selected model has prediction errors.

Our simulated *actual* PUE measurements at each temperature step come from a data set provided by the MGHPCC. We aggregate samples of PUE measurements by 5°F bins of wet bulb temperature, selecting the 5th percentile, median, and 95th percentile PUE measurements from each group. We run simulations of the data center where the bidding policies are not aware of the actual PUE. As a result, the $\bar{P}$ and R bids are unable to reflect the resulting increase or decrease to power consumption, so all adjustments to the incorrect site-wide power estimate must be managed through scheduling and power capping. This evaluation emulates the range of power prediction error that may occur if the data center operates at different efficiency shortly after placing a bid (e.g., due to cooling system changes or weather forecast error).

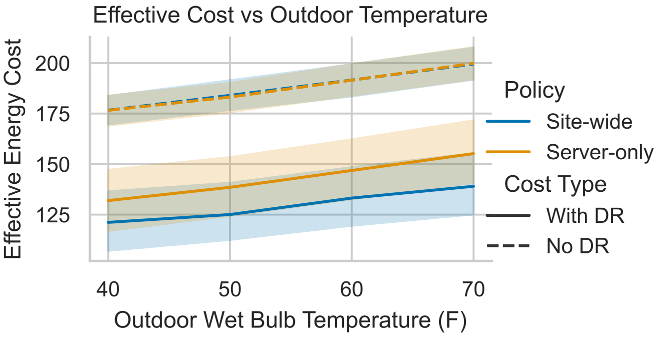

Fig. 3. Cost of demand response (DR) policies with and without site-wide awareness, compared to the cost of electricity purchase policies that do not participate in demand response programs, averaged across 3 simulations of the workload mixes described in Section IV. Shaded regions indicate the 95% confidence interval.

V. EVALUATION

In this section, we evaluate the results from our experiments in cost-saving opportunities under an ideal site-wide power model, as well as our experiments in the effects that site-wide power prediction errors have on a demand response policy.

A. Cost Savings

Fig. 3 shows the simulated costs of demand response policies with and without site-wide power models, compared

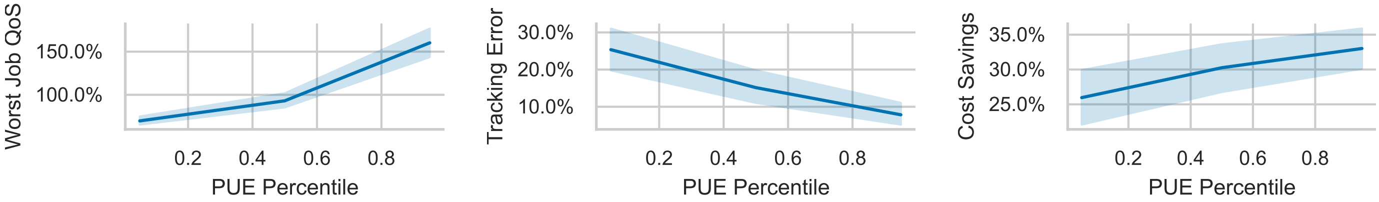

Fig. 4. Sensitivity of worst-case QoS, regulation signal tracking error, and cost savings as a response to corner cases of PUE modeling error. Shaded regions indicate the 95% confidence region across 3 trials of simulation over 4 temperature points and 5 workload mixes. *Cost savings* are relative to the cost of electricity without demand response. *Worst QoS* is the worst QoS degradation (relative to target QoS) across job types. *Mean tracking error* is the mean absolute error of the data center's actual power, with respect to the time-changing power target specified by the independent service operator.

with the costs of electricity-purchasing policies that do not take advantage of demand response programs. Regardless of the presence of demand response participation, costs trend upward with the simulated wet bulb temperature. That trend exists because more electricity is needed for non-server components, such as site-wide cooling systems, when outside temperature is higher.

The cost-saving opportunities of the site-wide demand response policies consistently outperform the opportunities of the server-only demand response policies. The average improvement is 1.3x savings in the site-wide policies, relative to the savings in the server-only policies.

B. Sensitivity Analysis

We first evaluate the high-level trends in our cost and performance metrics when there is an error in side-wide power consumption estimation. Next, we look at the properties of the worst-performing scenario underneath those trends.

1) High-Level Trends: Fig. 4 shows the sensitivity of our demand response cost and performance metrics, as the AQA policy over-predicts or under-predicts the site-wide power consumption. Low PUE percentiles indicate scenarios where the PUE is less than the typical PUE values that are used in the site-wide power model, so they indicate cases where the bidding policy is likely to bid higher than its needed power. Similarly, the higher PUE percentiles indicate scenarios where the bidding policy is likely to bid lower than its needed power.

As the site's actual power consumption trends from over-bidding to under-bidding scenarios, the QoS of the worst-performing job increases in both mean and in variance. Both of these trends occur because it is easier to let jobs run at full speed when there is a larger power budget.

Although over-bidding scenarios meet QoS needs more easily, they do increase the upper range of requested power targets from the independent service operator will not be easily achievable in the data center. Because over-bidding cases are more difficult to achieve all requested power targets, we see that lower PUE percentiles cause greater tracking error in the demand response power management policy.

These experiments cover the case where $\bar{P}$ and R selection are based on incorrect estimates of the site's power properties, so tracking error is the only remaining cost component described in Section IV-A. This means that although cost savings increase with the site's actual PUE, that is simply because the tracking error decreases in that case, and it comes with the previously-discussed impact on QoS degradation.

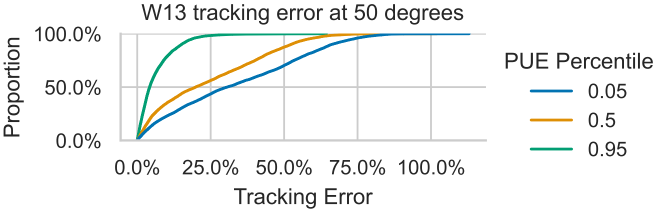

Fig. 5. Cumulative distribution function of target-power-tracking error as a percent of the R bid, for the W13 workload mix at 50°F wet-bulb temperature.

2) Worst-Performing Scenario: The worst-performing scenarios occur when the external wet-bulb temperature is near 50°F, where there is a transition between the data center's free cooling mode of operation and its active cooling mode of operation, as shown in Fig. 2. While the model indicates the PUE is 1.2 at 50°F, the 5th, 50th, and 95th percentiles of PUE at 50°F are 1.15, 1.17, and 1.26, respectively. In this section, we look more closely at behavior of the demand response policies on mix W13 at 50°F.

Workload mix W13 has a higher system utilization than the other workload mixes. This means that the data center has less capability to modulate its power level without causing work to accumulate in the scheduler's job queue. Fig. 5 shows the power-tracking error of the data center when applying the site-wide power management policy in different PUE prediction error scenarios. This case results in better tracking error when the site-wide power is under-estimated, and still struggles to track the target power near the median case. This suggests that the workload mix could benefit from standby jobs and preemptible jobs as described in the original AQA work [7].

In contrast to the more spread out tracking error distribution in the over-bidding case, the QoS degradation plots in Fig. 6 show more spread-out distributions in the under-bidding case. Furthermore, not all applications are impacted equally, as some have more drastic changes in QoS degradation across the PUE percentile cases.

The applications most sensitive to under-predicting the site power demand are `lu.D.112` and `mg.D.32`, which transition from far under threshold degradation in the 0.05 PUE percentile case to above threshold for 50% of runs in the 0.95 PUE percentile case. Both of those applications stay

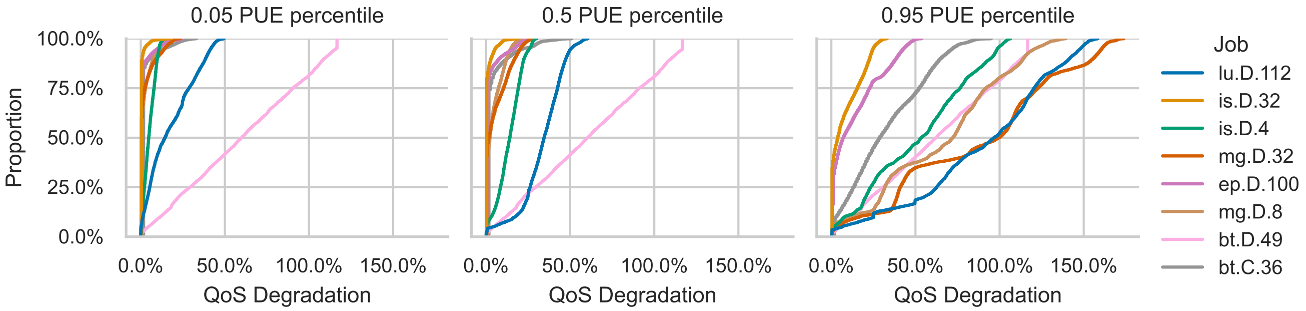

Fig. 6. Cumulative distribution functions of QoS degradation as a percent of each application's QoS threshold, for the W13 workload mix at 50°F wet-bulb temperature.

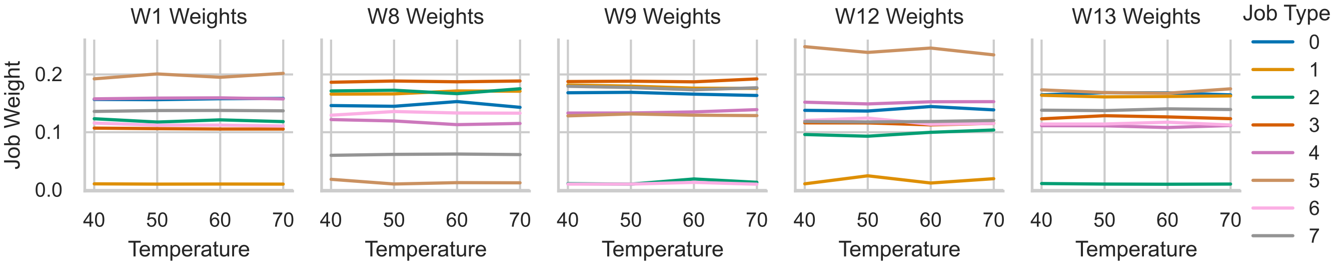

Fig. 7. Job weights selected by training the AQA demand response policy against each workload mix. Job types refer to the *App* column in table I. Job indices are presented here since they do not map to the same job names across workload mixes.

well under their QoS thresholds in all temperatures evaluated in the experiments with ideal site-wide power prediction (Section IV-A experiments).

C. Application Weights

Applications weights used by the AQA demand response policy are largely insensitive to the data center's site-level efficiency, as indicated by the nearly flat lines for all workload mixes in Fig. 7. This suggests that it is sufficient to train the AQA weights under a single simulated relationship between server power and site-wide power consumption for each workload mix. The learned weight could be used as an initial state in future searches for $\bar{P}$ and R bids across different PUE scenarios.

VI. Conclusion

Data centers globally consume a lot of power. As electricity markets work to adopt renewable portfolio standards, there will be increasing benefit to ensure big consumers are able to adapt with time-varying availability. Data centers that participate in demand response need accurate prediction and control of their power consumption, but they still need to commit to the QoS needs of their users.

This work proposes that adding a simple site-wide power model to existing QoS-aware demand response policies can improve the cost-saving opportunities while still being able to meet QoS constraints. We perform simulations of multiple mixes of applications on a 6000-server cluster with a simple PUE model for site-wide power consumption. We demonstrate that the presence of accurate PUE-based power predictions, we can seize 1.3x cost savings compared to QoS-aware demand response without site-wide power awareness.

Our evaluation of experimental results includes some observations that could be applied for future improvements to QoS-aware demand response management algorithms. Specifically, we note takeaways related to job weights used in the AQA and AQA-SW policy, and takeaways related to demand response risks from using a site-wide power model.

First, the job weights learned by the AQA policy's training mechanism depend on the relationship between server power properties and job performance properties, but not non-server power properties of the site. Future work may use this knowledge to reduce the search space when working with long-lived workload mixes across changing site-wide power efficiency.

Second, we note that there are competing risks in QoS and power tracking, from over-bidding or underbidding due to site-wide power-modelling inaccuracy. In our simple PUE model, the greatest risks are near the modeled data center's transition between two different modes of site-level cooling operation. Future work in site-wide power management policies could investigate methods to proactively adjust bids for reduced risk when operating in a low-confidence region of a site-wide power model.

Acknowledgment

We would like to thank the Massachusetts Green High Performance Computing Center for providing data and for their help understanding site-wide power components.

References

[1] "Renewable energy explained," https://www.eia.gov/energyexplained/renewable-sources/portfolio-standards.php, Jun. 2021.

[2] "International energy outlook 2021," https://www.eia.gov/outlooks/ieo/pdf/IEO2021_ChartLibrary_Electricity.pdf, Oct. 2021.

[3] "PJM manual 12: Balancing operations," https://pjm.com/~/media/documents/manuals/m12.ashx, Jun. 2022.

[4] B. Knowles, "ACM TechBrief: Computing and climate change," *ACM Technology Policy Council*, Nov. 2021.

[5] H. Chen, M. C. Caramanis, and A. K. Coskun, "The data center as a grid load stabilizer," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 105–112.

[6] A. Pahlevan, M. Zapater, A. K. Coskun, and D. Atienza, "ECOGreen: Electricity cost optimization for green datacenters in emerging power markets," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 2, pp. 289–305, 2021.

[7] 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.

[8] ——, "A data center demand response policy for real-world workload scenarios in hpc," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2021, pp. 282–287.

[9] "Efficiency - data centers - google," https://www.google.com/about/datacenters/efficiency/, 2021.

[10] Y. Fu, X. Han, K. Baker, and W. Zuo, "Assessments of data centers for provision of frequency regulation," *Applied Energy*, vol. 277, p. 115621, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920311247

[11] M. Maiterth, G. Koenig, K. Pedretti, S. Jana, N. Bates, A. Borghesi, D. Montoya, A. Bartolini, and M. Puzovic, "Energy and power aware job scheduling and resource management: Global survey — initial analysis," in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2018, pp. 685–693.

[12] T. Patki, N. Bates, G. Ghatikar, A. Clausen, S. Klingert, G. Abdulla, and M. Sheikhalishahi, "Supercomputing centers and electricity service providers: A geographically distributed perspective on demand management in europe and the united states," in *High Performance Computing*, J. M. Kunkel, P. Balaji, and J. Dongarra, Eds. Cham: Springer International Publishing, 2016, pp. 243–260.

[13] A. A. Chien, R. Wolski, and F. Yang, "The zero-carbon cloud: High-value, dispatchable demand for renewable power generators," *The Electricity Journal*, vol. 28, no. 8, pp. 110–118, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1040619015001931

[14] "MGHPCC," https://www.mghpcc.org/, 2021.