PowerMorph: QoS-Aware Server Power Reshaping for Data Center Regulation Service

ALI JAHANSHAHI, NANPENG YU, and DANIEL WONG, University of California, Riverside, USA

Adoption of renewable energy in power grids introduces stability challenges in regulating the operation frequency of the electricity grid. Thus, electrical grid operators call for provisioning of frequency regulation services from end-user customers, such as data centers, to help balance the power grid's stability by dynamically adjusting their energy consumption based on the power grid's need. As renewable energy adoption grows, the average reward price of frequency regulation services has become much higher than that of the electricity cost. Therefore, there is a great cost incentive for data centers to provide frequency regulation service.

Many existing techniques modulating data center power result in significant performance slowdown or provide a low amount of frequency regulation provision. We present PowerMorph, a tight QoS-aware data center power-reshaping framework, which enables commodity servers to provide practical frequency regulation service. The key behind PowerMorph is using "complementary workload" as an additional knob to modulate server power, which provides high provision capacity while satisfying tight QoS constraints of latency-critical workloads. We achieve up to 58% improvement to TCO under common conditions, and in certain cases can even completely eliminate the data center electricity bill and provide a net profit.

CCS Concepts: • **Hardware → Enterprise level and data centers power issues**; • **Software and its engineering → Power management**; • **Computer systems organization → Architectures;**

Additional Key Words and Phrases: Data center, power management, regulation service, quality of service, co-location

ACM Reference format:
Ali Jahanshahi, Nanpeng Yu, and Daniel Wong. 2022. PowerMorph: QoS-Aware Server Power Reshaping for Data Center Regulation Service. *ACM Trans. Archit. Code Optim.* 19, 3, Article 36 (August 2022), 27 pages.
https://doi.org/10.1145/3524129

1 INTRODUCTION

Environmental regulations and falling costs are driving the rapid adoption of renewable energy resources (e.g. wind and solar energy). During the past decade, electricity generation from wind energy has nearly tripled from 95,000 GWh to 254,000 GWh, and solar energy has grown nearly

This work was supported in part by NSF Grant CCF-1815643, California Energy Commission EPC-16-030, and the University of California, Riverside.
Authors' addresses: A. Jahanshahi and N. Yu, University of California, 459 Winston Chung Hall, University of California, Riverside, USA, 92521; emails: ajaha004@ucr.edu, nyu@ece.ucr.edu; D. Wong, University of California, Riverside, Winston Chung Hall, University of California, Riverside, Riverside, California, USA, 92521; email: danwong@ucr.edu.

1544-3566/2022/08-ART36
https://doi.org/10.1145/3524129

40x from 2,200 GWh to 81,000 GWh [29, 41]. Overall, the percentage of total generation due to renewable energy has increased from 12.3% to 19.7% [41]. However, the integration of renewable energy and its intermittent behaviour present challenges in maintaining electrical grid stability.

Electrical grids typically have an operating frequency of 50Hz (e.g. China, European countries) or 60Hz (e.g. United States, Canada). The operating frequency of the electrical grid can easily lose balance if the *supply* of electricity generation does not match the *demand* of electricity consumption. To combat this in modern smart grids, regional electric grid operators—also known as **Independent Service Operators (ISO)**—call for conventional power plants and end-use customers, such as data centers, to provision for *frequency regulation services*. The ISOs periodically send a *frequency regulation signal* to these entities which will accordingly adjust their electricity consumption/generation to help stabilize the grid frequency. In return, participants receive monetary benefits. With the increasing integration of solar and wind, and increasing grid instability, the price of frequency regulation has increased significantly [29], providing growing incentives along with opportunities for data center participation.

Conventional frequency regulation services are provided by electricity generators. However, generators tend to be slow in adjusting electricity generation and it is only feasible for larger, longer fluctuations in electrical grid conditions. More recently, batteries distributed across the electrical grid have been utilized for providing real-time frequency regulation services which require electricity adjustments every two seconds. However, using batteries suffer from poor battery lifetime due to the need to charge/discharge every two seconds and also the amount of regulation provisioned can fade if the battery is either fully charged or discharged [18, 35].

As an alternative, data centers have recently emerged as a compelling candidate for participation in frequency regulation services by providing significant regulation service provision and providing the ability to vary electricity consumption dynamically. Data centers consume 2% of US electricity usage, representing a large portion of overall electricity usage [22], providing a large potential source of regulation service provision. In the past, data centers have been explored to participate in various types of *demand response* including voluntary load reduction [24, 85], and peak shaving/power capping [14, 20, 31, 31, 89] through techniques including DVFS [14, 89], thread packing [12, 14], co-scheduling [31, 36, 70], and consolidating cores [6]. In addition, energy storage devices (i.e. batteries) can be used to achieve peak shaving by discharging during peak electricity usage periods and charging during low electricity usage periods [42, 48, 82]. However, relying on UPS batteries for peak shaving can result in shorter battery lifetime and jeopardizing power backup, and also requires significant capital expense investments.

Prior works [6, 8, 95] have attempted to adapt power capping techniques for participation in frequency regulation services and for load following [48]. These works target batch (best-effort) or HPC workloads, which typically run at maximum server power and are allowed to be slowed down (up to 200%) to track the regulation signal. Furthermore, they rely on existing power management knobs which limits the amount of regulation provision capacity that can be provided to the amount of power consumed by the workload.

Still, there are several significant limitations toward enabling *practical* frequency regulation services. First, data centers typically run a mix of latency-critical workloads and best-effort workloads. These prior techniques assume relaxed QoS targets and tolerated slowdowns of up to 200% QoS degradation [8, 95, 96], which would be intolerable for latency-critical applications. Second, it is unclear how incoming request traffic variability can be handled in concert with frequency regulation signals. Finally, the majority of prior works have been conducted through analytical models (at best, models derived from empirical measurements) which do not capture the real-world variability of latency-critical workloads [11, 39]. In this work, we present PowerMorph , the first work to *demonstrate support for data center frequency regulation in latency-critical environments*. The main

novelty of this work is that it is the first to achieve frequency regulation of servers running latency-sensitive workloads through the introduction of a novel knob (complementary workloads). While co-location of LC and BE workloads and throttling dummy load have been proposed in prior works, our work is the first to show how to carefully coordinate co-location, throttling of complementary workloads, and DVFS in order to maximize frequency regulation provisioning under ms-scale latency constraints. This work opens up frequency regulation to a whole new class of widely-used data center workloads/servers that previously was unattainable.

In this work, we make the following contributions:

- (Section 3). Identify the challenges of achieving frequency regulation service participation in commodity data centers running latency-critical workloads.
- (Section 4). We propose PowerMorph, a QoS-aware server power reshaping framework, enabling data centers to provide frequency regulation services using only computational resources under latency-critical data center conditions.
- (Section 5). We show that PowerMorph can accurately track frequency regulation signals in real-time and reshape server power profiles. In a small-scale data center evaluation, we observed total electricity cost savings of up to 71% and TCO ($ per throughput) improvement of up to 56% under common conditions. Under favorable conditions, which occur 10% of the time, it is even possible to completely eliminate electricity cost and achieve net profit. We also compare the total electricity cost of a data center providing frequency regulation using energy storage technique (Flywheel) and a cluster-level frequency regulation technique using CPU resource limiting and idle server modulation (EnergyQARE [8]) with PowerMorph.

2 BACKGROUND AND MOTIVATION

In this section, we will first provide an overview of frequency regulation service and the potential opportunities for electricity cost savings. Then we'll provide an overview of other common techniques used to optimize data center energy efficiency, such as power over-subscription and workload co-location. Then in the next section, we'll provide an overview of the limitations of existing work in providing practical data center frequency regulation service and motivate the need for PowerMorph.

2.1 Overview of Frequency Regulation Service

In order to maintain electrical grid stability, electrical grids must maintain operating frequency between 58.98Hz–60.02Hz in the United States. In traditional power grids, this is achieved by constantly adjusting the generator output to match the electricity consumption of consumers. However, as renewable energy sources such as wind and solar are integrated into the power grid, the intermittent nature of solar and wind causes significant variation in the electrical generator side. These sources have limited ability to adjust electrical generation supply in order to match consumer demand, and traditional power sources cannot vary power quickly enough to balance out the nature of solar and wind variation. Therefore, power system operators have recently begun allowing end-use customers to help maintain the electrical grid frequency.

Frequency Regulation Markets: In order to maintain the operating frequency of the electric grid at rated values, power system operators call for the provision of *frequency regulation services* from end-user customers and thermal power plants in day-ahead or real-time markets. The frequency regulation service *provision resources*, such as a data center, submit their estimated energy consumption baseline and frequency regulation service provision capability into the corresponding market either a day in advance (for day-ahead market), or an hour in advance (for real-time market). The energy consumption baseline is denoted as P_{avg} (i.e. the average amount of power the

data center is expected to consume in the next day/hour) and the amount of frequency regulation service is denoted by R (i.e. the amount of power the data center can vary on-demand).

For real-time markets, estimates can be made at the start of the hour at 60-minute granularity (e.g. energy consumed over the next hour) or at five-minute granularity (e.g. energy consumed for every five-minute interval over the next hour), depending on the ISO support. We find that making hour-ahead 60-minute granularity bids in the real-time market provides the ideal trade-off in forecasting accuracy, as it is difficult for data centers to forecast its usage a day ahead, or to estimate usage over the next hour in five-minute granularity. In this paper, we utilize PJM real-time market which only allows bids at 60-minute granularity.

Regulation signal: The frequency regulation service provision resources (e.g. a data center) have to modulate their power consumption to follow a *frequency regulation signal*, $r(t)$, which falls into the range of $[-1, 1]$. The frequency regulation service signal is broadcast *every two seconds* by the ISO based on the current state of the power grid. ISOs ensure that that the difference between two consecutive values of $r(t)$ does not exceed 0.5% of R [67], which means that the frequency regulation signal is relatively slow-moving compared to the variability experienced in servers. Examples of regulation signals can be seen in Figure 7.

By setting the energy consumption baseline (P_{avg}) and the amount of frequency regulation (R), the data center should keep its power consumption at time t to be $P_{avg} + r(t) \cdot R$. The energy charge of the data center at time period t equals to the product of P_{avg} and locational marginal price of energy at time t. The revenue (reward) that the data center receives at time period t from providing frequency regulation service equals the product of the amount of frequency regulation service (R) and the price of frequency regulation service price at time period.

Quantifying Quality of Frequency Regulation Service Provision: The revenue received from frequency regulation service is also dependent on, and proportional to, the *quality* of the provided regulation service. In other words, the magnitude of the revenue depends on how well a frequency regulation service provision resource (e.g. data center) can track the frequency regulation signal. The quality of tracking is quantified by a *performance score* [67]. In quantifying the performance score, the electricity market does not differentiate between the uncertainty of data center demand vs the inability to follow regulation demand. Performance score is calculated with Equation (1):

$$\text{Performance Score} = \frac{1}{3}(\text{Delay} + \text{Accuracy} + \text{Precision}) \tag{1}$$

Delay is the time delay between the frequency regulation signal and the point of its highest correlation with the regulation service provision resource's power consumption signal. *Accuracy* is the correlation or degree of relationship between the frequency regulation signal and the regulation resources' power consumption time series. *Precision* is calculated based on the instantaneous error between the regulation signal and the regulating resource's response.

ISOs typically certify a resource for regulation service provision after the resource achieves a performance score of 75% or better on three consecutive successful tests [67]. Once frequency regulation resources are qualified for regulation service provision, they have to maintain a performance score of 40% or higher, otherwise, they will be disqualified from future frequency regulation service provision [67].

Reward pricing and Electricity cost: Figure 1 shows the electricity cost (in \$/MWh) and the reward pricing (in \$/MWh) combination for a 1-year period in 2018 from PJM Interconnection [66]. The top and right histogram distribution shows the probability distribution function of electricity cost and reward pricing, respectively, in order to show the density of the scatter plot. Based on this, we can observe that electricity cost is typically in the \$20–\$40 per MWh range and the reward pricing is typically in the \$30–\$100 per MWh range. The diagonal lines in the scatter plot represent

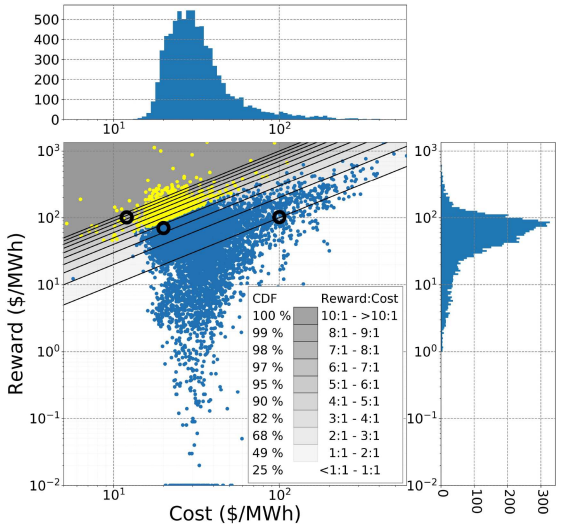

Fig. 1. Electricity cost and corresponding Regulation reward pricing in 2018 [66].

the reward to cost ratio, with the lowest line representing price parity. 75% of the time we observe reward pricing greater than or equal to electricity cost. Therefore, ample opportunities exist for data centers to take advantage of favorable reward pricing. In fact, ***we observed that there are certain reward to cost ratios that result in net profit***, as highlighted by the yellow-colored dots. That is, at reward to cost ratios above 4, we observe that 10% of the time[1] the reward revenue from frequency regulation completely offsets the electricity cost resulting in overall profit. With the increasing adoption of renewable energy, it is expected that the reward to cost ratio will only become more favorable [29]. Clearly, there is a great financial incentive for data center participation in frequency regulation markets.

Regulation Service vs Reducing Power Consumption: Data center operators try to optimize different aspects of the data centers to reduce costs and maximize monetary benefits as long as the optimization does not violate the **Service Level Agreement (SLA)**. Lowering data center energy consumption is of great importance because electricity costs are a major operation expense in data centers. To reduce the data center energy costs, numerous approaches have been proposed to minimize the energy consumption of servers [11, 39, 54, 56, 80, 87], or decrease the server's peak power without violating the SLA [3, 31, 71, 89].

Counter-intuitively, we show that regulation service mechanisms can enable data centers to reap monetary benefits *without* the goal of minimizing server power consumption. As shown in Figure 1, the reward to cost ratio is commonly 2x–10x. Due to these reward to cost ratios, there may be greater monetary benefits to participating in frequency regulation service than to minimize server power consumption—*in many cases it may be beneficial to have the server consume more power.*

2.2 Overview of Server Co-location

The traditional data center technique to improve the energy efficiency of data centers revolved around increasing the utilization of existing power infrastructure and servers. Typically, many servers run at lower utilization, and therefore consume less power than its nameplate power [20].

[1] Assumes a data center at moderate 40% load and 80% performance score.

This is especially true of servers running latency-critical workloads which typically exhibit request-response patterns where its utilization depends on the amount of incoming requests. A common technique to improve the energy efficiency of data center servers is to increase the utilization of the servers. Since servers commonly are lowly utilized [20], co-locating many jobs can significantly improve server utilization. For example, server virtualization is a commonly used technique to allow co-location on a single hardware server.

More recently, there has been significant work done in exploring the safe co-location of common data center workloads, such as best-effort batch-type workloads and latency-critical workloads. Many works exist in supporting safe co-location of latency-critical and batch workloads to increase server utilization and scheduling of safe co-location pairs [9, 15, 16, 55, 57, 61, 62, 65, 72, 90, 91, 94]. For example, Heracles [57] dynamically manages multiple hardware and software isolation mechanisms to ensure that latency-critical workloads meet their strict QoS targets while maximizing the resource given to best-effort tasks. More recently, safe co-location works have explored how to enable co-location of multiple latency-critical workloads [61, 65] by quickly adjusting resource isolation in a fine-grain manner.

In our work, the goal is to provide practical data center frequency regulation for latency-critical data center workloads. Due to their low utilization, the amount of frequency regulation provision available is severely limited. In order to increase the amount of frequency regulation provision available, we aim to co-locate latency-critical workloads with a complementary best-effort workloads with standard commercially available isolation mechanisms. Incorporating more advanced co-location policies would enable even better isolation of latency-critical and best-effort workloads, resulting in better tail latency results.

2.3 Overview of Data Center Power Capping

A common technique to improve the utilization of power infrastructure is to over-subscribe the number of servers in the data center and then limit the power consumption to safe levels under power emergencies. Power emergencies can occur when the amount of power consumed by servers exceed the amount of power that can be provided by the data center. Typical techniques to handle these power emergencies are known as peak shaving or power capping. A lot of research has been conducted on power-capping across single server, clusters/data center, or combination of them.

Server-level power capping: Peak shaving (power capping) limits the peak power consumption either the data center- or server-level resulting in lower peak demand charge. Power capping can be achieved with a wide range of techniques, which leverage computational resources. These techniques include DVFS [14, 89], thread packing [14], CPUJailing [37], co-scheduling of power-complementary workloads [31], consolidating cores [6], and using batteries [1, 27, 42, 82].

Data center-level power management: Most power management techniques at data center rely on meticulous coordination of server-level power capping techniques [20, 31, 37, 50, 84, 88] in conjunction with leveraging **power distribution units (PDUs)** [51, 73, 92].

Supporting power capping has some similarities to supporting frequency regulation. For example, both require the data center (or server) power consumption to meet a certain power level. In the case of power capping, this power level is a static power level, while in frequency regulation this power level is time-varying based on the regulation signal. However, there exists a critical distinction that presents unique challenges for frequency regulation. In power capping, the nominal utilization and power consumption level is at a maximal level and power capping techniques aim to *decrease* the power consumption through various means (i.e. DVFS, resource limiting, etc.). In the case of frequency regulation, the data center power level must be able to *decrease* or *increase*, depending on the regulation signal. In addition, the capacity for power increase/decrease must

be significant enough to provide a sufficient level of frequency regulation provisioning in order to obtain sufficient reward. In comparison to power capping techniques, PowerMorph not only requires servers to reduce power, but also follow and increase power.

3 CHALLENGES TOWARD PRACTICAL DATA CENTER FREQUENCY REGULATION UNDER LATENCY-CRITICAL CONSTRAINTS

Due to the large electrical load that data centers consume, data centers make a good candidate for participation in regulation services. Prior works have investigated the challenges and benefits of incorporating data centers into power grids as regulation resources [54] for demand response and frequency regulation service. However, most work focuses on the electricity market mechanisms on how to incentivize data centers to participate [54] or explores potential benefits through extensive modeling [30, 49]. However, these prior works do not adequately demonstrate practical implementations, and their challenges, for realizing data center participation in frequency regulation services. In our work, we specifically address how data center frequency regulation can be supported under more realistic environments which run latency-critical workloads. In this section, we will highlight the challenges towards achieving practical data center frequency regulation with commodity servers and our approach to overcome it.

How to maximize regulation provision? A key challenge that latency-critical workloads present is that servers tend to be lowly utilized due to the request-response nature of the workload [2]. This presents a challenge since the lower utilization of latency-critical workloads limits the amount of frequency regulation provision that can be provided. This contrasts to best-effort batch workloads which tend to run at near maximum utilization and provide a readily available large dynamic power range to modulate power.

Another key challenge to maximizing the amount of regulation service provision is the need to provide *symmetric* frequency regulation. While many power modulation techniques, such as DVFS and core shutdown, can already provide symmetric frequency regulation, their provision amount can be limited. For example, if a server commits a total of 20W for regulation service, then it must be able to either increase (up to $P_{avg} + 20$) or decrease (down to $P_{avg} - 20$) power consumption as requested. However, certain scenarios can lead to violations. For example, if core sleep states are used to reshape power and the server utilization is low, then there may not be enough cores to put to sleep to regulate the power down to satisfy the regulation signal which negatively affects the *performance score*. To maintain quality regulation performance, only a limited amount of power can be provisioned for frequency regulation.

To address these limitations and to maximize regulation provision, we pair the latency-critical workload with a co-located complementary workload to provide *offset power* which symmetrically increases the amount of room to modulate power up and down.

How to practically support complementary workloads? A key contribution of PowerMorph is the use of complementary workloads to regulate server power. Essentially, we co-locate a best-effort workload that we can modulate. Utilizing complementary workloads presents several challenges. Specifically, the complementary workload needs to be able to handle the high variability of the latency-critical workload and needs to avoid performance-degrading contention with the latency-critical workload. Due to the request-response nature of the latency-critical workload, server utilization tends to exhibit high short-term variability and is prone to bursty behavior [11]. This presents a unique challenge for the complementary workload as it needs to modulate its utilization to complement the latency-critical workload and at the same time aim to accurately track the moving regulation signal.

If not carefully co-located, the complementary workload may also contend with the latency-critical workload causing QoS degradation. As shown previously, there exists a large body of work

Table 1. Overview of Limitations of Existing Works Enabling Data Centers to Provide Regulation Service

Power Modulation Techniques	Server/ Cluster	Workload Support[1]	QoS Criteria	Workload Service Time Constraint	RS Provision
DVFS [49]	Server	Sim.	No QoS Support	ms	Low
Forced idle injection [59]	Server	BE	No QoS Support	s	Medium
CPU resource limit [6]	Server	BE	BE *sojourn* time[2]	s	Medium
Power Capping, Job sched. [96]	Cluster	Sim.	BE *sojourn* time	min/hour	Medium
CPU res. lim., Idle server [8]	Cluster	BE	BE *sojourn* time	s	Medium
RE, EES, VM Allocation[3] [63]	Cluster	Sim.	BE *sojourn* time	min	High
RAPL, Job sched./Queue [95]	Cluster	BE	BE *sojourn* time	s	High
Dummy load, DVFS [83]	Cluster	Sim	No QoS Support	ms	High
Complementary Workload, DVFS[4]	Cluster	LC&BE	LC tail latency	ms	High

[1]BE: Best-effort/Batch, LC: Latency-critical, Sim: Simulation || [2]*sojourn* time = queue time + execution time.
[3]RE: Renewble Energy, EES: Electrical Energy Storage || [4]This work: PowerMorph.
PowerMorph (the last row) enables data centers with co-located batch and latency-critical applications (very tight QoS constraint) to participate in regulation service to reduce the data center electricity costs.

that propose co-location frameworks to allow latency-critical and best-effort workloads to safely co-locate. Although co-location frameworks that support multiple latency-critical workloads exist, we do not consider co-locating multiple latency-critical workloads as a complementary workload since the strict QoS requirement of the latency-critical workloads would eliminate any possible power modulation opportunity. In order to maintain the tight QoS of the main workload, the goal of this work is to answer "*What level of isolation is required to safely co-locate complementary workloads with latency-critical workloads for regulation service?*" and also to see "*How does co-located workloads variance impact regulation service quality?*"

How to reshape power? A major challenge of data center frequency regulation is the selection of techniques to modulate power to track the regulation signal. The challenge here is the time granularity of the regulation signal requires the data center to vary its power every two seconds and the need to provide sufficient and symmetric regulation provision. We mainly focus on servers since they consume the largest portion of the total data center power [38, 86]. Furthermore, servers provide a large dynamic power range for providing regulation service. Therefore, these computational resources are a large source of untapped regulation service provision that does not require the capital expense overheads of utilizing energy storage devices (i.e. flywheels, batteries)[2] and are readily available in commodity data centers. Within servers, by far the largest consumer of power is the processor, followed by main memory [38, 86]. Memory tends to not be significantly energy proportional as main memory has significant static power due to the need for DRAM refresh [58, 74]. Processors, on the other hand, are extremely energy proportional due to aggressive low power states such as idle power states (power gating) and dynamic voltage frequency scaling, which makes them an ideal candidate.

Table 1 shows a list of common techniques that can modulate data center power and their limitations. At the cluster-level, power can be potentially reshaped by migrating load in order to consolidate workloads to a subset of active servers and turn off idle servers. In addition, idle servers can be turned on/off such that the idle power can act as a form of power modulation. However, load migration takes in the order of seconds or minutes, and turning idle servers on/off can take in the order of seconds, both of which are not responsive enough to track regulation signals. Load can also be modulated by queuing up jobs that are going into the cluster. While potentially more responsive, this approach can result in significant delays in job processing time. In many cases, these techniques can tolerate and enforce QoS targets with up to 200% performance degradation.

[2]See Section 5 for frequency regulation comparison against Flywheel.

Due to this high tolerance, standby jobs are able to be used while delaying requests is not tolerable for latency-critical workloads.

Therefore, in order to modulate cluster-level power, we would require coordination with server-level techniques which are more responsive. Potential knobs here involve DVFS [49] and core sleep states [59] which can be modulated in the order of milliseconds. However, DVFS can only reshape dynamic power, which limits the amount of regulation provision that it can provide. Core sleep states can provide more benefits by also taking advantage of static power. Techniques such as CPU resource limits [6] can be combined with DVFS and core sleep. Hardware power limiting mechanisms, such as RAPL [40], provide power capping through hardware-controlled DVFS. Many of these server-level techniques are coordinated with cluster-level techniques [8, 95] to provide higher levels of frequency regulation provisions. However, a major limitation of these techniques is that they can significantly slow down the running workload, which is detrimental in latency-critical environments.

To solve these aforementioned limitations, we introduce using *complementary workload* where we utilize a co-located application as a knob for power reshaping. By modulating a complementary workload, we can provide millisecond power reshaping (to mask the high variability of the latency-critical workload and meet the granularity of the regulation signal), provide a high provision for frequency regulation (to both increase and decrease power consumption), and can meet tight QoS targets.

How to coordinate cluster-level power reshaping? Another challenge of data center frequency regulation is in coordinating regulation service across all servers in the data center in order to maximize cost benefits. Cluster-level coordination occurs at two time-scales. Every hour the data center has to make a bid for the amount of frequency regulation (R). Every two seconds the data center as a whole has to follow a regulation signal. This two-second regulation signal does not provide ample opportunity for complex cluster-level optimizations. On top of that, the data center can have various cluster scheduling policies (such as load-balanced or consolidated) which can interfere with cluster-level coordination of frequency regulation.

Therefore, we propose a hierarchical approach where servers are allocated individual frequency regulation provisions and enforced locally to meet the timing requirements of the regulation signal, and regulation provisions are reallocated every hour globally which enables more time-intensive optimization policies to maximize cost benefits. We found this hierarchical approach provides good provisions and adapts to various cluster scheduling policies.

4 POWERMORPH

The goal of PowerMorph is to provide practical data center-wide frequency regulation using commodity servers. The PowerMorph framework coordinates server power reshaping using DVFS and complementary workload provides performance isolation to maintain tight QoS, and maximizes rewards by providing symmetrical regulation service provision. Intuitively, PowerMorph utilizes complementary workload to add offset power to maximize the amount of regulation service provision. PowerMorph dynamically adapts to the power behavior of different types of applications running on the server resulting in more flexibility and robustness. Figure 2 demonstrates the server- and date center-level components of PowerMorph. In this section, we describe how server-level components of PowerMorph work, then expand the proposed server-level regulation service approach to enable data center participation in regulation service.

4.1 PowerMorph Profiler

Targeting data centers with commodity servers in this paper, each server has a specific power consumption pattern based on its hardware resources and the workload running on it. In order to

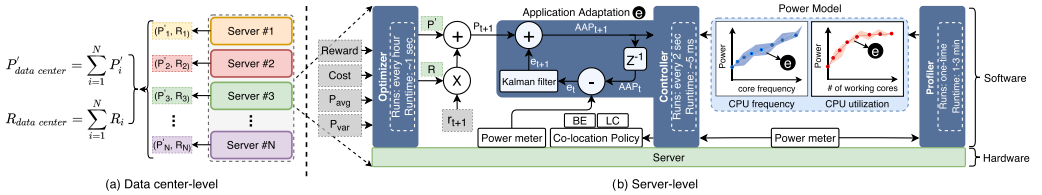

Fig. 2. PowerMorph overview. (a) Hierarchical coordination of servers for data center-level regulation service. (b) Server-level components of PowerMorph. Grey dashed boxes are PowerMorph inputs. Profiler runs on each server only one time to obtain the power model of the server. Optimizer runs every hour to determine if the server participates in RS. R and P_{off} are computed by Optimizer. Since the profiled power model can vary for different applications, we adapt the computed power to the application's power behavior. **Application adapted power (AAP)** is the target power of the server to be set in the next two-second time step which is the adaption of the power model interpolation error (e in black circle) for currently running applications on the server. The controller runs every two seconds to control the co-location policy module.

control a server's power, i.e. providing regulation service, PowerMorph requires the power consumption pattern of the server which we call *Power model*. Profiler is run once on each server to sample and capture the *Power model* of the server with a workload running on it. Depending on the granularity of frequency scaling that the server's hardware supports, the profiling operation takes about one-to-three minutes.

We use deep learning training workload on the server since they are both computation- and memory-intensive. The captured power model is built by interpolating the samples of (utilization, frequency) pairs each of which corresponds to a P_{CPU} and P_{DRAM}. Due to interpolation and using one workload type to profile the power consumption pattern of a server introduces an error to the power model, ⓔ in Figure 2(b). We use a 1D Kalman filter to make PowerMorph capable of adapting to different workloads which will be explained in Section 4.3.3.

4.2 PowerMorph Optimizer

The profit of providing regulation service depends on the average power usage of the server which is determined by the workload (P_{avg}) and the regulation provision (R) that the server is able to provide. Optimizer is responsible for picking a (P_{avg}) and (R) that maximize the data center profit.

4.2.1 Maximizing Regulation Provision with Offset Power. Using *power range lines*, Figure 3 gives an illustrative overview of how PowerMorph adjusts the server's CPU cores to reshape its power, while avoiding impacting the latency-critical workload adversely. The circle markers on the power range lines are points of interest for the server's power when running its target workload. P_{min} and P_{max} represent the server's minimum (active idle) and maximum power consumption. P_{avg} is the average power consumption of the latency-critical workload running on the server (with no participation in regulation service). Due to natural variations in workload load, there is a natural variance in the power consumption of the server at a given utilization. This variance range is represented by the smaller solid circles. In this illustrative example, we assume we have 16 cores on the server, where the workload on the server can be serviced by packing all of its work in the first five cores. Therefore, on average, the amount of power consumed is due to the first five cores. This represents a case where the workloads' utilization is typically ~30%, but can vary from ~20%–40% due to real-world short-term variation. We note that all cores typically do not consume the same amount of power as the utilization-power curve is non-linear.

In order to maximize the amount of regulation service provision, we need to provide a large symmetrical range. *Our approach is to introduce offset power (through the use of complementary*

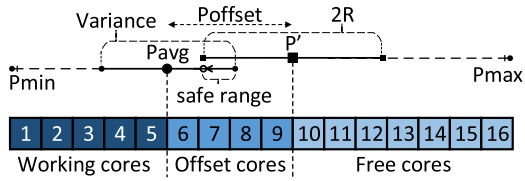

Fig. 3. PowerMorph processor core organization and operation.

workload) so that we have a larger dynamic power range to utilize. In the figure, P_{offset} is the offset power added to P_{avg}. Therefore, the effective average power of the server is $P' = P_{avg} + P_{offset}$. $2R$ is the regulation server provision that is available. Therefore, the server's power can range from $P' - R$ to $P' + R$, represented by the smaller square markers. These regulation service parameters are readjusted by Optimizer every 60 minutes depending on the regulation service market, data center workload, and workload variance.

4.2.2 Determining Regulation Provision and Offset Power. PowerMorph tries to minimize the total electricity cost by picking a proper offsetted power P' (reported to ISO as the average power) and $|R|$, for any given workload (i.e. P_{avg} and P_{var}), reward (rew), and electricity cost ($cost$). Equation (2) shows the optimization formula solved by PowerMorph.

$$\begin{aligned} minimize : Total\ elec.\ cost &= P'_{cost} - Reward \\ st : P'_{cost} &= P' \times cost \\ Reward &= |R| \times rew \\ P_{avg} + \frac{P_{var}}{2} &< P' < P_{max} \\ Total\ elec.\ cost &< threshold \times P_{avg} \times cost \end{aligned} \tag{2}$$

Recall that $[-R, +R]$ has to be symmetrical around P'. Therefore, for any given P', $|R|$ is calculated as follows:

$$|R| = min\left(P_{max} - P', P' - \left(P_{avg} + \frac{P_{var}}{2}\right) + safe\ range\right) \tag{3}$$

As P' increases, $|R|$ increases up to a point, then begins to decrease (because $+R$ eventually becomes restricted by P_{max}). Similarly, as P' gets close to $P_{avg} + \frac{P_{var}}{2}$, it becomes restricted by the safe range (shown as an arrow with a hollow circle in Figure 3) which represents the limit of lowering frequency while safely meeting QoS.

To find the optimal P', we solve the optimization formulated in Equation (2) with exhaustive search (as illustrated in Figure 4) by gradually increasing P_{offset} from $P_{avg} + \frac{P_{var}}{2}$ to P_{max}, which we call sweeping. For every P', we estimate the total monetary benefit of participating in regulation service and select the combination that maximizes the benefit. *This optimization runs every hour* when the data center bids how much regulation service provision it can provide. Based on our experiments, this step takes under a second. Therefore, this algorithm has negligible overheads.

In order to pick R and P', we need to know the server's P_{avg} and its variance power (P_{var}). Much research has been done on predicting these parameters for data centers based on their historical load traces [4, 5, 17, 21, 53, 81, 93]. In this work, our aim is not in proposing new load prediction algorithms for data centers. Instead, we can rely on these prior works to be able to predict the average load of the data center, which we can then use to estimate the server's P_{avg} and P_{var}. We evaluate the impact of power prediction inaccuracy in Section 5.

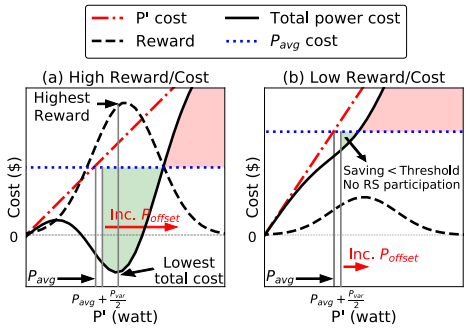

Fig. 4. Illustrative example of selecting P' for two extremes. Savings are shaded in green (or net profit if less than 0), and red represents increased cost if we participate in RS.

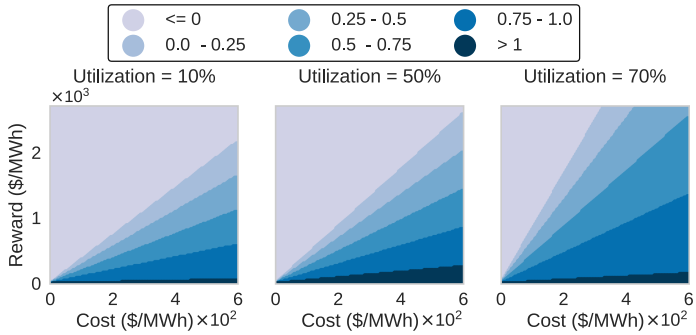

Fig. 5. Design space exploration of benefits (Normalized total electricity cost).

When to participate: Figure 4 shows illustrative examples of our algorithm in picking P'. Figure 4(a) shows a case with high reward/cost ratio. The dotted line shows the P_{avg} cost of the server without participating in RS. By increasing P_{offset}, reward ($|R|$) first increases, and then decreases (dashed line). Meanwhile, by increasing P_{offset}, the electricity cost increases (dash-dotted line). Since reward/cost is high, monetary reward outweighs the electricity cost introduced by P_{offset}, resulting in savings (green shaded region). However, at some point (after the "Highest Reward" point), due to shrinking $|R|$, the total electricity cost begins to increase to the point that it exceeds the P_{avg} cost (dotted line), i.e. red area. In other words, after some point, it is not beneficial to increase P_{offset} anymore. PowerMorph picks P' that minimizes total electricity cost (solid line).

Figure 4(b) shows an example in which reward/cost is low. Since reward is low and electricity price is high, electricity cost savings is only observed with small P_{offset} before electricity cost overheads dominate. As P' increase, total cost quickly exceeds the P_{avg} cost (dotted line), i.e. red area. In such scenarios, in which the green area is very small, an even small misprediction of either P_{avg} or P_{var} leads to losing money. To avoid such scenarios, PowerMorph uses a *threshold* to be conservative in participating in regulation service. If the minimum total electricity cost (for the best P') is higher than $threshold * (P_{avg}cost)$, PowerMorph decides not to participate in RS. We use $threshold = 0.95$ in our experiments.

The impact of reward and electricity cost: We also performed a space exploration to investigate how the total electricity cost is affected by reward to cost ratio and P_{avg}. Figure 5 shows total electricity cost normalized based on the average server power (P_{avg}) for all possible reward prices and electricity cost values. We ran the experiment for the scenarios in which the server is

running with 10, 50, and 70% utilization each of which corresponds to a P_{avg}. In this experiment we assume there is no workload variation, and PowerMorph is able to follow the regulation signal with a performance score of 80%.

In Figure 5, normalized total electricity cost equal to 1 means the total electricity cost of the server participating in regulation service is equal to the electricity cost of the same server without participating in regulation service. There is a point at which the reward (R) outweighs the electricity cost introduced by P_{offset}, the area at which normalized total electricity cost is greater than 1.

PowerMorph withdraws from regulation service for all points that have normalized total electricity cost of greater than 1, which is the dark blue-colored area in Figure 5. The lower the normalized total electricity cost, the more monetary benefit we get. Negative normalized total electricity cost (less than zero) means not only we do not pay for the electricity we use, but we also earn money at that point, which is shown by grayish color in Figure 5. In Figure 5, the area at which we earn money (negative normalized total electricity cost) shrinks as the server utilization (P_{avg}) increases. The reason is that at higher utilization regions, the amount of R that we can provide starts to decrease, and as a result we do not get a large benefit.

4.3 PowerMorph Controller

To provide regulation service, the server power needs to be adjusted every two seconds. *PowerMorph Controller* calculated the target power of the server (P_{t+1}) based on the regulation signal ($r(t)$), regulation provision (R) and $P^{'}$ calculated by *PowerMorph Optimizer*, as well as issuing the proper commands to the co-location policy module.

4.3.1 Core Organization. To provide isolation between the complementary workload and the latency-critical workload to maintain tight QoS, we pin tasks to specific groups of cores. Based on real-time utilization of the latency-critical workload, the *Working core* set is pinned with the latency-critical workload, and other resources, such as cache, are allocated to them to meet their target QoS. The power consumption of the working cores is P_{avg} with some power variance, due to workload variance.

The *Offset cores* are dynamically assigned between either the latency-critical workload or the complementary workload as the server's utilization varies. These cores increase the server's power consumption in order to provide symmetry to increase or decrease server power, as well as to increase the amount of regulation service provisioning that we can provide. The *Free core* set is used to increase target power by adjusting the complementary workload when needed. By organizing PowerMorph into three core types, we can provide different functionalities for regulation service, along with the server's original latency-critical workload, in a way that decouples the performance impact with reshaping server power.

The number of cores assigned to the latency-critical workload is readjusted in real-time to dynamically provide performance isolation. If latency-critical workload needs more computation resources, an offset core (or free core if no offset cores exist) is reallocated and converted into a working core instantly. Then, other cores are reevaluated in a way that the server's power follows the regulation signal.

Due to variation in server load, the offset cores also act as cores that absorb this noise and minimize core reallocation events. In order to remove the switching overhead, we added hysteresis to the switching. Switching an offset core to a working core occurs instantly when the workload requires more core. On the other hand, if a working core has not been used for a while, we convert it into either an offset core or free core, whichever can preserve P'. Adding this hysteresis makes the isolation more robust, reliable, and has almost no overhead.

4.3.2 Complementary Workload. By artificially inflating the utilization of the server by P_{offset}, we can essentially follow regulation signals solely by scaling the utilization of the complementary workload and varying power around P'. Fundamentally, there is a trade-off between how much power we offset by (extra electricity cost) and how much reward we get by increasing our regulation service provision (more regulation reward). Supporting utilization scaling requires us to explore 1) what type of complementary workloads to use, and 2) how to control the complementary workload.

Best-effort complementary workload selection: A common approach to improve the energy efficiency of data centers is to co-locate best-effort workloads with latency-critical workloads in order to increase the utilization of the servers. To select a best-effort complementary workload, we assume the server can rely on a multitude of prior works that select safe co-location workload pairs [9, 55, 57, 61, 62, 65, 91].

Complementary workload isolation: One of our goals is to *identify the level of isolation required to safely co-locate* complementary workload with latency-critical workloads. Towards this end, we evaluate using isolation mechanisms that are readily available in commercial off-the-shelf servers. While more sophisticated workload co-location mechanisms exist [61, 62, 91], our evaluation is conservative and would obtain even better results with more advanced techniques.

To provide isolation and preserve QoS in co-location scenarios with best-effort workloads, we follow a similar scheme to Heracles [57]. Since latency-critical workload has priority over the best-effort workload, we continuously monitor the resource requirement latency-critical workload and adjust hardware resources allocated to that. Using `taskset` command, we pin the latency-critical and best-effort workloads to separate cores so that there is no interference between them. In order to help latency-critical workload run faster, we increase the priority of its processes using `nice` command. To isolate shared resources such as LLC, we utilize Intel's Cache Allocation Technology [34] which allows partitioning of cache between tasks. Currently, no memory bandwidth isolation techniques exist. In [57], memory bandwidth availability was maintained by scaling down the number of BE cores. In our experiments, we observed that the main sources of contention come from core and cache contention, and memory contention has a minuscule effect.

Controlling utilization of best-effort workloads: One challenge of using best-effort complementary workloads is that we cannot direct the best-effort workloads to limit utilization directly. Therefore, to limit the utilization (and the power consumed by the server), we need to throttle these workloads' utilization using existing Linux system tools. We observed `taskset` achieves a better performance score as this method is more robust to noise introduced by variation.

4.3.3 Morphing Server Power. Power morphing is guided by a sampled profiled power model that interpolates the power curves shown in Figure 6. For each (utilization, frequency) pair, we have P_{CPU} and P_{DRAM}, separately. Every cycle (two seconds) we need to determine the target power ($P_{t+1} = r_{t+1} * R + P'$) based on the new regulation signal r_{t+1}, the chosen R, and determine if we should increase/decrease power.

Mapping target power to target utilization/frequency: Utilization (on offset/free cores) and frequency scaling (on working cores) are the knobs we use to morph server power. To achieve a target power, we need to select a utilization/frequency point given our current operating point (as illustrated in Figure 6).

To decrease power, PowerMorph first removes free/offset cores allocated to best-effort complementary workload (❸→❷). If still necessary, PowerMorph further decreases the frequency of the working cores while remaining within the safe range (❷→❶). To increase power, PowerMorph first increases the frequency of cores allocated to the working cores (❶→❷). Next, PowerMorph increases the server power by allocating offset/free cores to the best-effort complementary workload until the target power is reached (❷→❸).

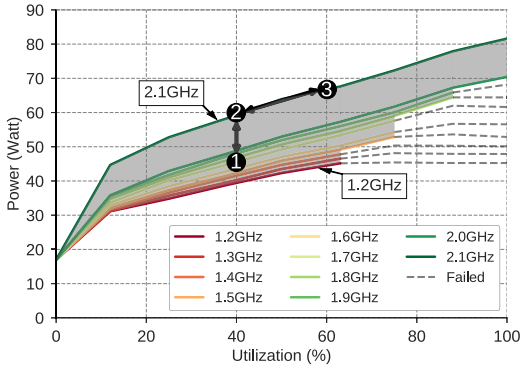

Fig. 6. Utilization - Power profile of our experimental server. Dashed line conditions where QoS fails. Solid lines represent where QoS is satisfied. The overall "safe range" where QoS passes are highlighted in grey.

Adapting to noise and application types: Due to noise/error introduced by inaccuracies to the profiled power model, non-deterministic nature of real systems, and application type-dependent power consumption pattern, setting the server utilization/frequency may not lead to the target power. To address this problem, a *1D Kalman filter* is integrated into PowerMorph . The noise/error of the previous cycles (e_t) is fed into the filter to get an estimate of the error (e_{t+1}) we predict for the power model to have for the next two-second interval. Adding e_{t+1} to the target power (P_{t+1}) we will have the *Application adapted power* (AAP_{t+1}) for the next interval. Then, using the power model obtained by the *Profiler*, PowerMorph maps AAP_{t+1} to utilization/frequency which is going to be set.

The adaptive capability of PowerMorph provides high-quality regulation service with different application types, i.e. memory- or compute-intensive. For example, for memory-intensive applications, the extra memory power (compared to the memory power in the profiled power model) is inputted as noise/error to the filter. Therefore, the application adjusted power (AAP_{t+1}) would be less than the target power (P_{t+1}) to alleviate the adverse effect of extra memory power usage. We also noted that depending on the workload, the maximum server power varies while the shape of the power curve remains similar. To account for this difference, we derive a scaling factor that scales the profiled power curve to the running workload's power at a given utilization when making power predictions.

Maintaining QoS: In order to maintain QoS under varying loads and regulation signals, we monitor the amount of latency slack (the difference between observed latency and target tail latency) at run-time. If the observed latency approaches the target tail latency, then we would need to increase the amount of latency slack available to avoid any QoS violations. The way to do this is by lowering the amount of best-effort workload that is co-locating. By opting to maintain the latency slack, we essentially trade-off the performance score (and amount of reward we can obtain) to ensure QoS levels are met.

4.4 Data Center-level Regulation Service

As illustrated in Figure 2(a), to provide frequency regulation across the data center, we utilize a hierarchical approach where each server is allocated its own responsibility of frequency regulation provision. For example, Server A (due to its workload or hardware resources) can provide 10W for frequency regulation service and Server B can provide 20W for frequency regulation service. The regulation signal is then broadcast to every server, where every server is responsible for tracking a regulation signal with respect to their own regulation provision.

Table 2. Workload Utilization Trace Properties

Trace name	Avg. load (%)	Variance	Min (%)	Max (%)
email	10.38	10.5	3	34
msg-store	32.1	10.77	21	59
high-util	50.5	15.32	25	75

email and *msg-store1* are from [87].

To support this, we reallocate regulation provision (R) responsibility every hour. Every hour, we broadcast the reward and electricity pricing to each server and each individual server will determine its average power ($P^{'}$) during the one-hour interval as well as the amount of frequency regulation provision (R) it can provide. The server's regulation provision will then be aggregated at the data center-level and sent to the ISO. To determine the data center's estimated power consumption, we aggregate the estimated power of all running servers. Since this reallocation occurs once every hour, this process can utilize more complex optimization.

5 EVALUATION

Platform setup, tools, and benchmarks: We run all experiments on a small-scale data center of six servers with an Intel Xeon E5-2620 v4 processor, which has 16 physical cores, 128GB of DDR4 DRAM. Power of the server is sampled through Intel PCM [33]. The *Web Search* benchmark from *CloudSuite* [64] is used as a representative latency-critical workload. The target tail latency was selected as the 95th percentile tail latency of Web Search running in isolation. To obtain the target tail latency, we adopt the same methodology established in prior works [11, 91]. We obtain the target tail latency at the "knee" of the utilization[3]-tail latency curve, where queues and tail latency begins to grow—which we observe to occur at ∼90% of the maximum supported RPS.

Workload utilization traces: We evaluate Web Search under realistic varying workload utilization traces from Table 2. We use two workload traces of differing variance (*email*, and *msg-store1*) from [87]. These traces were collected from institutional data centers representing a wide range of workloads including web serving, email services, and data stores.

According to [2], the utilization of servers running latency-critical workloads is typically around 20-40% of max RPS. However, to evaluate PowerMorph at higher utilization ranges, we used a synthetic high utilization load (*high-util*). We also evaluate a scenario where the cluster has *mixed* workload utilization where each trace is run by two servers.

Best-effort complementary workloads: When selecting a complementary workload, it is imperative that this workloaddoes not degrade the QoS of the latency-critical workload. Complementary workloads can simply be safe co-located best-effort workloads in co-located data centers. While outside the scope of this work, we assume that safe co-location workload pairs can be assigned dynamically to the servers from a multitude of prior works on identifying safe co-location pairs at run-time [9, 55, 57, 61, 62, 65, 91]. To select a candidate complementary workload in our experiments, we evaluated a range of applications from SPEC2017, PARSEC3.0, and machine learning training (AlexNet, VGG, LeNet) built on Keras. We observed that all of these best-effort workloads can safely co-locate with our target latency-critical workload using existing isolation mechanisms available in commodity servers. For our complimentary workload, we selected AlexNet training. ML workloads give us a throughput metric (training epochs per second) [32] which we can use to quantify throughput and TCO impacts.

[3]Our metric for *utilization* is with respect to the maximum achievable request-per-second of the LC workload and not the OS reported CPU utilization (i.e. as reported in `top`).

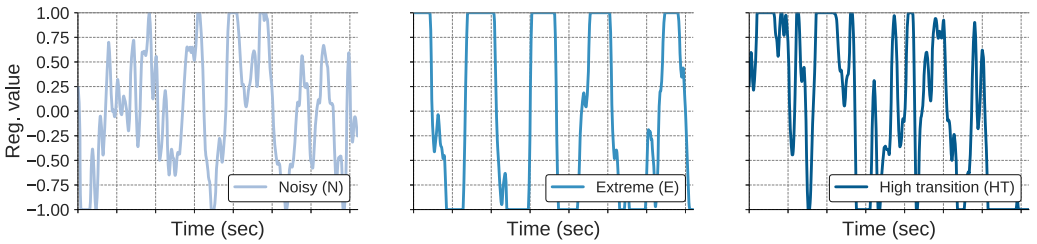

Fig. 7. Regulation signals used to evaluate PowerMorph.

Regulation signal selection: We select three regulation signals from PJM regulation signal archive [68] selected from 2018. Figure 7 shows the regulation signals we chose for evaluating PowerMorph . Since we are participating in hour-ahead regulation market, we picked one hour slices. We chose `Extreme (E)` with regulation signal that stays in the highest and the lowest power points for extended periods of time. We chose `High Transition (HT)` which have frequent min-to-max power change requests. We select `Noisy (N)` to evaluate how accurate PowerMorph can track small changes in the regulation signal.

Regulation reward and Electricity cost selection: The regulation reward and electricity cost is broadcast every hour (for hour-ahead regulation market). We selected three pairs of (regulation rewards, electricity cost) shown with hollow black circles in Figure 1. (70, 20) is selected from the area with the highest density (the most common scenario). (101, 12) represents a high reward/cost ratio pair. (102, 100) represent a reward/cost ratio that is approximately 1 where electricity cost is high. *Note that with high electricity pricing, reward price is typically high and of similar magnitude.* For ratios where price is greater than reward, PowerMorph typically decides not to participate in regulation service.

Evaluation scenarios: In our evaluation, we consider the following scenarios. `LC + BE` represents a baseline scenario where **best-effort (BE)** workloads are co-located to increase the utilization and efficiency of the servers. `LC + BE + RS` represents the co-location case that is participating in regulation service where the best-effort complementary workload is being regulated by PowerMorph.

5.1 Comparative Results

Figure 8 shows a comparative design-space exploration of various frequency regulation techniques across a range of reward-to-cost ratios. This figure runs every technique with our three workload utilization traces. We define total electricity cost = (cost of electricity consumption)–(reward obtained from regulation service) + (capital expense cost). Capital expense only applies to the Flywheel scenario.

Comparison to traditional data center-level energy-saving approaches: To reduce the data center energy costs, numerous approaches have been proposed to minimize the energy consumption of servers [11, 39, 54, 56, 80, 87], decrease the server's peak power without violating SLA [3, 31, 71, 89], or consolidate servers to turn off idle servers [13, 23, 52, 69, 77–79, 86]. Data center-level scheduling policies typically fall into two broad categories: *Uniform* load balanced and *Right-sizing*, which consolidates workloads in order to save power. In Figure 8, the electricity cost savings due to right-sizing is shown with the black horizontal lines and are normalized to each workload's electricity cost using Uniform scheduling. The *mixed* scenario is omitted for figure clarity. As workload utilization decreases, this results in more power-saving opportunities for right-sizing, with *email* resulting in ∼80% electricity cost savings.

We evaluate PowerMorph on top of both Uniform and Right-sizing scheduling. In the case of PowerMorph + right-sizing, the idle servers are not shut off to save power, but instead used

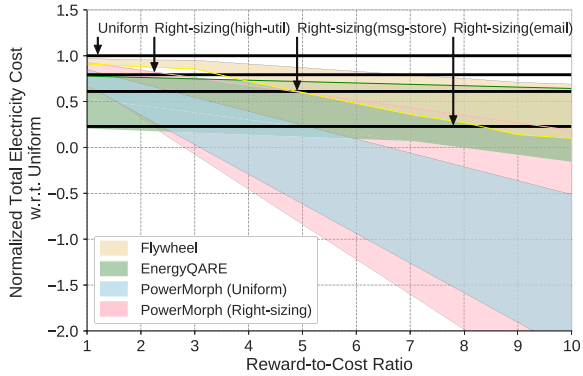

Fig. 8. Total electricity cost of cluster compared to energy storage technique (Flywheel) and a cluster-level frequency regulation technique using CPU resource limiting and idle server modulation (EnergyQARE [8]). The colored region represents a range of cost under various utilization traces. All results normalized to Uniform load balanced scheduling.

entirely by the complementary workload to provide regulation service. For scenarios where reward-cost ratio is above 3 (a common scenario), PowerMorph consistently saves more in electricity cost compared to right-sizing. Despite PowerMorph consuming more power by not shutting down idle servers, the amount of reward far outweighs the cost of increased power consumption. In certain cases, PowerMorph even provides net electrical cost profit where the amount of reward exceeds the electricity consumption cost! *Counter-intuitively, we show that regulation service mechanisms can enable data centers to reap monetary benefits without the goal of minimizing server power consumption.*

Comparison with Flywheel energy storage system: We compare against Flywheel [60], a data center-level energy storage system that has been shown to be one of the best suited for frequency regulation applications [7]. Energy storage devices facilitate frequency regulation service by either charging or discharging to change the data center's power consumption profile without impacting the underlying workload. However, energy storage devices incur high upfront capital cost expenses. In our small-scale experiment, we provision the Flywheel to be similar to the peak power consumption of our cluster with capital expense cost of $2,400/KW spread over 20 years and power-energy ratio of 0.25 which is typical of commercial products today [60]. Overall, we found that Flywheel is effective and can save up to ~90% of the total electricity cost with reward-to-cost ratio of 10. However, we found that the capital expense of the Flywheel can significantly reduce the overall monetary benefit. PowerMorph , by comparison, can provide significant regulation provision without any upfront capital cost, leveraging the available power flexibility in servers.

Comparison with alternative data center-level frequency regulation technique: We compare against EnergyQARE [8] which runs on top of right-sizing scheduling policies and coordinates server-level CPU resource limiting with turning idle servers on/off for additional regulation provision capacity. Even though EnergyQARE can enforce QoS targets of up to 200% slowdown, the amount of monetary benefits is limited (averages ~80% savings) due to the relatively smaller capacity of regulation provision that CPU resource limiting and idle servers can provide.

5.2 PowerMorph Evaluation Results

As shown in Figure 8, PowerMorph consistently outperforms alternative techniques for data center frequency regulation. When running on top of Right-sizing, PowerMorph is more sensitive to utilization load as the number of idle servers fluctuate, and hence, the amount of regulation

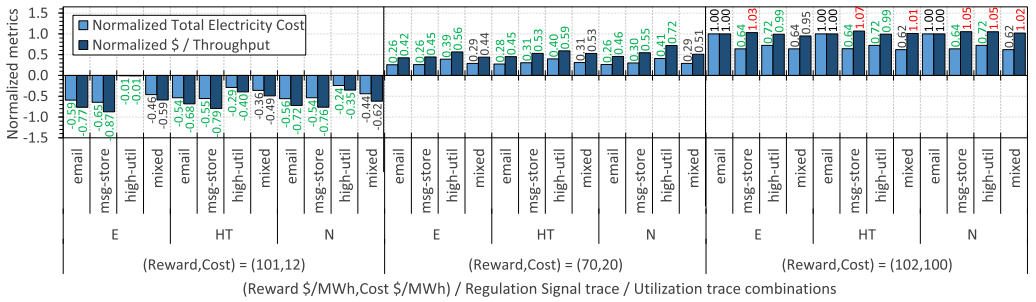

Fig. 9. Total Electricity Cost and Cost per Throughput for LC+BE+RS (PowerMorph) normalized to LC+BE. Bars of 1.00 represents scenarios that decides to not participate in RS. In *all* scenarios, PowerMorph improves total electricity cost.

provision. Running on top of Uniform scheduling is more challenging for PowerMorph as every server has our complementary workload co-located with a latency-critical workload which introduces more workload variance. Towards this end, the remainder of this evaluation focuses on PowerMorph on top of Uniform scheduling which is more challenging.

Figure 9 shows our experimental results for total electricity cost and total cost of ownership. The figure shows the result of the co-location case with regulation service (LC + BE + RS) normalized to the baseline co-location case (LC + BE). These scenarios are evaluated against various (reward, cost) conditions, regulation signal patterns, and workload utilization patterns as discussed previously. For certain scenarios, PowerMorph determines that it is not worth it to participate in regulation service; these are indicated when both total electricity cost and $/Throughput are both 1.0. We note that due to PowerMorph's hierarchical approach to cluster-wide coordination, we observed similar results as we scale across different server counts and hence our result is representative of larger clusters.

5.2.1 Total Electricity Cost. In the common case of (70, 20), PowerMorph can save 59%–74% of the total electricity cost. Even when the reward-cost ratio is not favorable, (102, 100), LC + BE + RS can still save 28%–38% of total electricity cost when participating in frequency regulation. For favorable cases (101, 12), we observed that the amount of monetary reward can outweigh the total electricity cost. In these scenarios, we observed that ***we can earn a net profit equivalent to up to 65% of the original total electricity cost!***

In general, *we observe that total electricity cost savings remain relatively stable across different regulation signals,* thus demonstrating that PowerMorph is able to efficiently handle arbitrary regulation signals.

5.2.2 Total Cost of Ownership. In order to estimate the impact of PowerMorph on total cost of ownership, we evaluate the Dollar amount spent on electricity per throughput ($/throughput). This gives us a more holistic evaluation metric that incorporates both throughput impact and electricity cost to evaluate if the throughput reduction of best-effort workloads justifies the gains in frequency regulation service participation.

Measuring TCO: To capture impact to the total cost of ownership, we evaluate $ per throughput. Note that typically the metric throughput per $ is used; however, due to having negative electricity cost this metric becomes difficult to understand. We simply take the inverse to represent TCO. This metric can simply be understood as the cost (or reward) for every unit of throughput the data center provides.

Table 3. QoS (Tail Latency) of LC Workload Normalized to the Target Tail Latency as Well as BE Workload QoS (Throughput) Normalized to Baseline Co-location Case for Different Utilization Traces in PowerMorph

Trace name	Normalized LC tail latency			Normalized BE throughput		
	(101, 12)	(70, 20)	(102, 100)	(101, 12)	(70, 20)	(102, 100)
email	0.52	0.64	0.68	0.55	0.19	1.0[1]
msg-store	0.52	0.63	0.52	0.43	0.14	0.22
high-util	0.49	0.52	0.52	0.45	0.29	0.43

[1]In this case, PowerMorph decides not to participate in RS.

Measuring throughput: For best-effort workloads, we use the number of training epochs per minute as the throughput metric. For latency-critical workloads, we use queries per second as the throughput metric. In order to quantify these two throughput metrics into a single metric, we use the **System Throughput (STP)** metric [19] which is commonly used to capture throughput in multiprogram environments. STP quantifies the total system throughput as follows:

$$STP_{server} = \frac{Throughput_{LCw/RS}}{Throughput_{LC}} + \frac{Throughput_{BEw/RS}}{Throughput_{BE}}$$

For a single server, ideal STP is equivalent to 2 since we're running two workloads (LC + BE). Values less than 2 indicate overall throughput decrease. The throughput in the denominator is the throughput when running the baseline co-location, while the numerator is the throughput when participating in regulation service. To quantify STP for a data center cluster, we simply take the summation of each server's STP where ideal STP is two times the number of servers.

TCO results: For the typical case (70,20), we observe TCO improvements of 28–58%. For favorable reward-cost ratio (101,12), we are now basically earning money for every unit of computational throughput. In this scenario, we are earning up to 87%, per unit of throughput, of what we would have paid for electricity cost per throughput unit.

For scenarios where reward-cost ratio is not favorable, (102,100), the $/throughput is around parity ranging from 1.03 to 1.15. The throughput decrease is mainly from the complementary workload and is due to PowerMorph deciding that the additional electricity cost of offset power does not out-weight the reward benefit of providing a larger regulation provision. Therefore, PowerMorph decides to participate with less offset power (and thus, less complementary workload). Even with a worse case $/throughput decrease of 15%, we still save 31% of total electricity cost. Therefore, system designers will need to carefully identify whether total electricity cost is more important or throughput is more important when running in these reward-cost range.

5.2.3 Quality-of-Service. QoS has been defined as the *sojourn* time of BE workload in previous works [8, 63, 96]. In this work, however, the QoS is defined as the latency of LC workload. Table 3 shows the average normalized tail latency of PowerMorph across different (reward, cost) conditions and individual utilization traces. Across all scenarios, not only is PowerMorph able to maintain QoS levels but also the QoS tail latency has been improved. Since PowerMorph regulates the utilization of the complementary workload to follow a regulation signal, we will always introduce less interference compared to the baseline co-location case. The co-location techniques utilized by PowerMorph are not the strictest which shows by utilizing more advanced co-location techniques, PowerMorph is capable of performing even better. Therefore, there is a large room to isolate workloads even more and obtain more profit in regulation service. Table 3 also shows the BE throughput (QoS) normalized to that of baseline co-location case.

Table 4. Average Performance Score of Providing Regulation Service by PowerMorph for Different Scenarios

Regulation Signal			Trace			Overall
E	HT	N	email	msg-store	high-util	Average
85.02	83.62	80.52	81.53	83.62	84.01	83.05

Although the QoS of BE workload is not considered in the PowerMorph optimizer, the result shows that the QoS degradation is about 60% on average and within the range of 45–86% when PowerMorph participates in RS which still meets the QoS limit defined in previous works which allow up to 200% QoS degradation [8, 63, 96]. According to [8], 200% QoS degradation is translated to 0.33 throughput degradation. As shown in Table 3, for some scenarios, PowerMorph is not able to keep BE QoS within the range reported by prior works.

5.2.4 Performance Score. Table 4 shows the average performance score of providing regulation service for different scenarios. Across all scenarios, PowerMorph is able to provide performance scores of >80 with an overall average of 83.05. Of all the regulation signals, Noisy signal is the hardest to track due to the need to track small changes in regulation signal. Even in this scenario, PowerMorph is able to obtain a performance score of 80.52. We observe that as the number of servers in the cluster increases, the overall data center performance score improves due to variation across servers having a masking effect of under-performing individual servers.

5.2.5 Impact of Average Power/Variation Misprediction. One of the goals of this paper is to investigate how co-located workload variance impacts regulation service quality. To investigate this, we artificially inject variation errors (misprediction) of −10, −5, +5, and +10W for one scenario. Figure 10 shows the impact of artificially injecting misprediction errors when predicting workload variation for *msg-store* and Noisy regulation signal. We find that performance score is not greatly impacted by variation misprediction, but normalized TCO is impacted slightly; no more than 5% difference in the worst case.

5.3 Mixed Workload Cluster

Figure 11 shows total electricity cost and total cost of ownership of a six-server cluster with combinations of workloads described in Table 2. Overall, we observe similar trends at the data center-scale similar to that of the single server scenario, thus, highlighting the feasibility of scaling out PowerMorph across the data center. In favorable cases, we observe profit of up to 46% of the total electricity cost. In the common case (70, 20) we save up to 71% of total electricity cost with 56% improvement to TCO. In the non-favorable cases, we achieve up to 37% improvement to total electricity cost with near parity TCO.

6 DISCUSSION

Frequency regulation, public vs private data centers: PowerMorph utilizes the server's performance metrics to maintain the QoS for the latency-critical application. Since workload performance metrics are required, this work assumes private data centers where the applications' requirement is known. Many previous works on data center frequency regulation [8, 63, 95, 96] use *sojourn* time as performance metric to measure the QoS of the batch workload which also is not practical in a public data center and is limited to private data centers.

Aggressively provisioned data centers: Without frequency regulation participation, the data center would aggressively provision and safely co-locate latency-critical and best-effort workloads. In this baseline case, the best-effort workload would run unconstrained. If we participate in

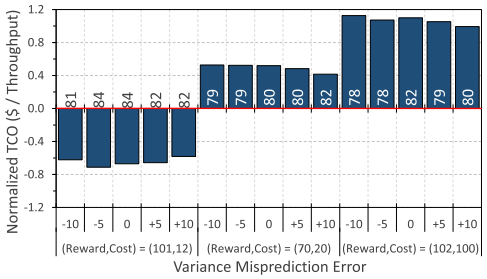

Fig. 10. Effect of workload variance misprediction on Normalized TCO and Performance score (labels on bars) for msg-store trace and noisy regulation signal N. 0 indicates no variance misprediction.

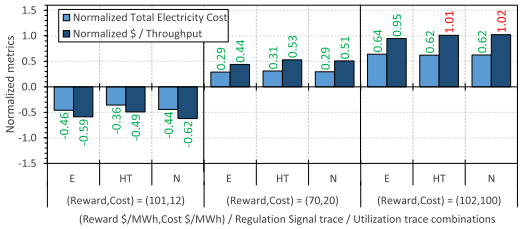

Fig. 11. Normalized Total Electricity Cost and Cost per Throughput of mixed workload utilization traces (2 servers per trace) normalized to LC+BE.

frequency regulation, PowerMorph will utilize the co-located workload to modulate power. This means that under frequency regulation the co-located workload will always be consuming less power (to track the regulation signal) compared to the baseline case. If there is a workload burst, we would handle this scenario similar to the baseline case by throttling the best-effort workload and prioritizing the latency-critical workload. However, aggressively provisioned data centers operate at higher utilization and have less power headroom, which can potentially limit the amount of frequency regulation provision that PowerMorph can provide in order to maintain availability.

Security concerns: Power attacks can create power emergencies that threaten the availability of aggressively-provisioned data centers [46]. In general, data center frequency regulation techniques are susceptible to such power attacks which can impact workload performance and overall cost returns. Power attacks can be detected based on attack features, feature extraction, or abnormal user behavior [10]. However, the attacker can evade this by changing the attack patterns and even attack the data centers with power attack detectors. PowerMorph can potentially provide a ground truth for power attack detection. For any given average power of the server (P') and $r(t)$, at any given time, the target power can be calculated and monitored by an automated system. The moment the power of a server does not follow the expected target power, it can be a sign of power attack which can be further investigated by more complex power attack detection methods.

Currently, PowerMorph relies on the workload average load and its variance which can be manipulated by attackers. We assume the data center is not compromised and it is secured from power or DDoS attacks which interfere with the predicted workload behavior of the data center leading to power consumption misbehavior. PowerMorph is most suitable for private data centers which have more control over the security. Also, PowerMorph framework runs on each server independent from the other servers in the data center, resulting in more security isolation in case attackers manage to compromise a small portion of servers.

7 RELATED WORK

The most relevant work in providing frequency regulation service in data centers was discussed previously in Section 3.

Renewable energy-powered data center: This intermittent nature of renewable energy poses many workload scheduling problems [25, 26, 75] and scheduling/design of power sources [26, 44, 45]. A major problem is *load matching*, where there is a need to balance the load power demand and local/global power generation. Load matching has been proposed at the processor-level [43, 47] by

using DVFS to tune load, by using stored energy devices [27, 28], and by coordinating local power generators to track power and power shaving to trim load demand [48]. These prior techniques mainly target batch workloads without tight millisecond-level QoS requirements, and also load matching at 15-minute granularities. In contrast, frequency regulation requires power readjustment every 2s and PowerMorph maintains ms-level QoS requirements. Due to this, PowerMorph can also be applied to load matching of renewable energy data centers, but not vice versa.

Batteries for RS: Leveraging UPS has also been considered to enable data centers participating in RS [30, 76] reduce the electricity costs of data center. While UPS can be leveraged to participate in regulation service, they incur significant capital expense and they are mainly designed for backup power, and not for the charge and discharge cycles required for regulation service which leads to lifetime issues.

8 CONCLUSION

In this work, we have proposed PowerMorph, a QoS-aware server-level power-reshaping framework which enables data centers to participate in regulation service by dynamically adjusting the servers' power consumption, providing us with up to 71% savings in electricity costs and up to 58% TCO improvement in common conditions. To the best of our knowledge, PowerMorph is the first practical demonstration of frequency regulation service under realistic latency-critical data center environments.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions.

REFERENCES

[1] Baris Aksanli and Tajana Rosing. 2014. Providing regulation services and managing data center peak power budgets. In *Proceedings of the Conference on Design, Automation & Test in Europe (DATE'14)*. European Design and Automation Association, 3001 Leuven, Belgium, Belgium, 143:1–143:4. http://dl.acm.org/citation.cfm?id=2616606.2616782.

[2] Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. 2018. The datacenter as a computer: Designing warehouse-scale machines, Third edition. *Synthesis Lectures on Computer Architecture* 13, 3 (2018), i–189.

[3] A. A. Bhattacharya, D. Culler, A. Kansal, S. Govindan, and S. Sankar. 2012. The need for speed and stability in data center power capping. In *2012 International Green Computing Conference (IGCC)*. 1–10. https://doi.org/10.1109/IGCC.2012.6322253

[4] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya. 2015. Workload prediction using ARIMA model and its impact on cloud applications' QoS. *IEEE Transactions on Cloud Computing* 3, 4 (Oct. 2015), 449–458. https://doi.org/10.1109/TCC.2014.2350475

[5] Katja Cetinski and Matjaz B. Juric. 2015. AME-WPC: Advanced model for efficient workload prediction in the cloud. *Journal of Network and Computer Applications* 55 (Sept. 2015), 191–201. https://doi.org/10.1016/j.jnca.2015.06.001

[6] Hao Chen, Can Hankendi, Michael C. Caramanis, and Ayse K. Coskun. 2013. Dynamic server power capping for enabling data center participation in power markets. In *Proceedings of the International Conference on Computer-Aided Design (ICCAD'13)*. IEEE Press, Piscataway, NJ, USA, 122–129.

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

[8] Hao Chen, Yijia Zhang, Michael C. Caramanis, and Ayse K. Coskun. 2019. EnergyQARE: QoS-Aware data center participation in smart grid regulation service reserve provision. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 4, 1 (Jan. 2019), 2:1–2:31. https://doi.org/10.1145/3243172

[9] Shuang Chen, Christina Delimitrou, and José F. Martínez. 2019. PARTIES: QoS-aware resource partitioning for multiple interactive services. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 107–120.

[10] Shenglei Chen, Dongyang Ou, Congfeng Jiang, Jing Shen, Li Yan, and Shuangshuang Guo. 2020. Power attack and detection technology in data centers: A survey. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 1–5.

[11] C. Chou, L. N. Bhuyan, and D. Wong. 2019. μDPM: Dynamic power management for the microsecond era. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 120–132.
[12] Marcus Chow, Kiran Ranganath, Robert Lerias, Mika Shanela Carodan, and Daniel Wong. 2021. Energy efficient task graph execution using compute unit masking in GPUs. In *Redefining Scalability for Diversely Heterogeneous Architectures Workshop (RSDHA)*.
[13] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, and Andrew Warfield. 2005. Live migration of virtual machines. *USENIX Association* (2005).
[14] Ryan Cochran, Can Hankendi, Ayse K. Coskun, and Sherief Reda. Pack & cap: Adaptive DVFS and thread packing under power caps. In *44th Annual IEEE/ACM International Symposium on Microarchitecture*.
[15] Christina Delimitrou and Christos Kozyrakis. 2013. Paragon: QoS-aware scheduling for heterogeneous datacenters. In *18th International Conference on Architectural Support for Programming Languages and Operating Systems*.
[16] Christina Delimitrou and Christos Kozyrakis. 2014. Quasar: Resource-efficient and QoS-aware cluster management. In *19th International Conference on Architectural Support for Programming Languages and Operating Systems*.
[17] Sheng Di, Derrick Kondo, and Walfredo Cirne. 2012. Host load prediction in a Google compute cloud with a Bayesian model. In *International Conference on High Performance Computing, Networking, Storage and Analysis*.
[18] EnergyStorageNews. 2022. PJM's Frequency Regulation Rule Changes Causing 'Significant and Detrimental Harm'. https://www.energy-storage.news/news/pjms-frequency-regulation-rule-changes-causing-significant-and-detrimental.
[19] S. Eyerman and L. Eeckhout. 2008. System-level performance metrics for multiprogram workloads. *IEEE Micro* 28, 3 (May 2008), 42–53.
[20] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. 2007. Power provisioning for a warehouse-sized computer. In *34th Annual International Symposium on Computer Architecture*.
[21] W. Fang, Z. Lu, J. Wu, and Z. Cao. 2012. RPPS: A novel resource prediction and provisioning scheme in cloud data center. In *2012 IEEE Ninth International Conference on Services Computing*. 609–616. https://doi.org/10.1109/SCC.2012.47
[22] ForesterNetwork. 2022. Data Center Demand Response. https://www.foresternetwork.com/distributed-energy/article/13036367/data-center-demand-response.
[23] Anshul Gandhi, Mor Harchol-Balter, Ram Raghunathan, and Michael A. Kozuch. 2012. Autoscale: Dynamic, robust capacity management for multi-tier data centers. *ACM Transactions on Computer Systems (TOCS)* 30, 4 (2012), 1–26.
[24] M. Ghamkhari and H. Mohsenian-Rad. 2012. Data centers to offer ancillary services. In *3rd International Conference on Smart Grid Communications*.
[25] Íñigo Goiri, Kien Le, Md. E. Haque, Ryan Beauchea, Thu D. Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. 2011. Greenslot: Scheduling energy consumption in green datacenters. In *International Conference for High Performance Computing, Networking, Storage and Analysis*.
[26] Íñigo Goiri, William Katsak, Kien Le, Thu D. Nguyen, and Ricardo Bianchini. 2013. Parasol and GreenSwitch: Managing datacenters powered by renewable energy. *SIGPLAN Not.* 48, 4, 51–64. https://doi.org/10.1145/2499368.2451123
[27] Sriram Govindan, Anand Sivasubramaniam, and Bhuvan Urgaonkar. 2011. Benefits and limitations of tapping into stored energy for datacenters. In *Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA'11)*. ACM, New York, NY, USA, 341–352.
[28] Sriram Govindan, Di Wang, Anand Sivasubramaniam, and Bhuvan Urgaonkar. 2012. Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters. In *17th International Conference on Architectural Support for Programming Languages and Operating Systems*.
[29] GreenTechMedia. 2022. In California, Solar and Wind Boost the Price of Frequency Regulation. http://www.greentechmedia.com/articles/read/in-california-solar-and-wind-boosts-the-price-for-frequency-regulation.
[30] R. Guruprasad, P. Murali, D. Krishnaswamy, and S. Kalyanaraman. 2017. Coupling a small battery with a datacenter for frequency regulation. In *2017 IEEE Power Energy Society General Meeting*. 1–5.
[31] Chang-Hong Hsu, Qingyuan Deng, Jason Mars, and Lingjia Tang. 2018. SmoothOperator: Reducing power fragmentation and improving power utilization in large-scale datacenters. In *23rd International Conference on Architectural Support for Programming Languages and Operating Systems*.
[32] Hamid Reza Imani, Jeff Anderson, and Tarek El-Ghazawi. 2022. iSample: Intelligent client sampling in federated learning. In *6th International Conference on Fog and Edge Computing (ICFEC)*.
[33] Intel. 2022. Intel Performance Counter Monitor. http://www.intel.com/software/pcm.
[34] Intel. 2022. Introduction to Cache Allocation Technology in the Intel Xeon Processor E5 v4 Family. https://software.intel.com/en-us/articles/introduction-to-cache-allocation-technology.
[35] IsoNewsWire. 2022. Redesigned Regulation Market Now in Effect. http://isonewswire.com/updates/2015/4/7/redesigned-regulation-market-now-in-effect.html.

[36] Ali Jahanshahi, Hadi Zamani Sabzi, Chester Lau, and Daniel Wong. 2020. GPU-NEST: Characterizing energy efficiency of multi-GPU inference servers. *IEEE Computer Architecture Letters* 19, 2 (2020), 139–142.

[37] Kostis Kaffes, Dragos Sbirlea, Yiyan Lin, David Lo, and Christos Kozyrakis. 2020. Leveraging application classes to save power in highly-utilized data centers. In *Proceedings of the 11th ACM Symposium on Cloud Computing*. 134–149.

[38] S. Kanev, K. Hazelwood, G. Wei, and D. Brooks. 2014. Tradeoffs between power management and tail latency in warehouse-scale applications. In *2014 IEEE International Symposium on Workload Characterization (IISWC)*.

[39] Harshad Kasture, Davide B. Bartolini, Nathan Beckmann, and Daniel Sanchez. 2015. Rubik: Fast analytical power management for latency-critical systems. In *Proceedings of the 48th International Symposium on Microarchitecture*. ACM, 598–610.

[40] Kashif Nizam Khan, Mikael Hirki, Tapio Niemi, Jukka K. Nurminen, and Zhonghong Ou. 2018. RAPL in action: Experiences in using RAPL for power measurements. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)* 3, 2 (2018), 1–26.

[41] Samuel Koebrich, Emily I. Chen, Thomas Bowen, Sydney Forrester, and Tian Tian. 2019. *2017 Renewable Energy Data Book: Including Data and Trends for Energy Storage and Electric Vehicles*. Technical Report. National Renewable Energy Lab. (NREL), Golden, CO (United States).

[42] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, and T. Simunic Rosing. 2012. Managing distributed UPS energy for effective power capping in data centers. In *2012 39th Annual International Symposium on Computer Architecture (ISCA)*. 488–499. https://doi.org/10.1109/ISCA.2012.6237042

[43] C. Li, X. Li, R. Wang, T. Li, N. Goswami, and D. Qian. 2013. Chameleon: Adapting throughput server to time-varying green power budget using online learning. In *International Symposium on Low Power Electronics and Design (ISLPED)*. 100–105. https://doi.org/10.1109/ISLPED.2013.6629274

[44] Chao Li, Amer Qouneh, and Tao Li. 2012. ISwitch: Coordinating and optimizing renewable energy powered server clusters. In *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA'12)*. IEEE Computer Society, USA, 512–523.

[45] Chao Li, Rui Wang, Tao Li, Depei Qian, and Jingling Yuan. 2014. Managing green datacenters powered by hybrid renewable energy systems. In *11th International Conference on Autonomic Computing (ICAC'14)*. USENIX Association, Philadelphia, PA, 261–272. http://www.usenix.org/conference/icac14/technical-sessions/presentation/li_chao.

[46] Chao Li, Zhenhua Wang, Xiaofeng Hou, Haopeng Chen, Xiaoyao Liang, and Minyi Guo. 2016. Power attack defense: Securing battery-backed data centers. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 493–505.

[47] C. Li, W. Zhang, C. Cho, and T. Li. 2011. SolarCore: Solar energy driven multi-core architecture power management. In *2011 IEEE 17th International Symposium on High Performance Computer Architecture*. 205–216. https://doi.org/10.1109/HPCA.2011.5749729

[48] C. Li, R. Zhou, and T. Li. 2013. Enabling distributed generation powered sustainable high-performance data center. In *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. 35–46. https://doi.org/10.1109/HPCA.2013.6522305

[49] Sen Li, M. Brocanelli, Wei Zhang, and X. Wang. 2013. Data center power control for frequency regulation. In *2013 IEEE Power Energy Society General Meeting*. 1–5.

[50] Shaohong Li, Xi Wang, Faria Kalim, Xiao Zhang, Sangeetha Abdu Jyothi, Karan Grover, Vasileios Kontorinis, Nina Narodytska, Owolabi Legunsen, Sreekumar Kodakara, et al. 2020. Thunderbolt:{Throughput-Optimized},{quality-of-service-aware} power capping at scale. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI'20)*. 1241–1255.

[51] Yang Li, Charles R. Lefurgy, Karthick Rajamani, Malcolm S. Allen-Ware, Guillermo J. Silva, Daniel D. Heimsoth, Saugata Ghose, and Onur Mutlu. 2019. A scalable priority-aware approach to managing data center server power. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 701–714.

[52] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew, and Eno Thereska. 2012. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Transactions on Networking* 21, 5 (2012), 1378–1391.

[53] Chunhong Liu, Chuanchang Liu, Yanlei Shang, Shiping Chen, Bo Cheng, and Junliang Chen. 2017. An adaptive prediction approach based on workload pattern discrimination in the cloud. *Journal of Network and Computer Applications* 80 (Feb. 2017), 35–44. https://doi.org/10.1016/j.jnca.2016.12.017

[54] Zhenhua Liu, Iris Liu, Steven Low, and Adam Wierman. 2014. Pricing data center demand response. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'14)*. ACM, New York, NY, USA, 111–123.

[55] Q. Llull, S. Fan, S. M. Zahedi, and B. C. Lee. 2017. Cooper: Task colocation with cooperative games. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 421–432.

[56] David Lo, Liqun Cheng, Rama Govindaraju, Luiz André Barroso, and Christos Kozyrakis. 2014. Towards energy proportionality for large-scale latency-critical workloads. In *Proceeding of the 41st Annual International Symposium on Computer Architecture (ISCA'14)*. IEEE Press, Piscataway, NJ, USA, 301–312. http://dl.acm.org/citation.cfm?id=2665671.2665718

[57] David Lo, Liqun Cheng, Rama Govindaraju, Parthasarathy Ranganathan, and Christos Kozyrakis. 2015. Heracles: Improving resource efficiency at scale. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA'15)*. ACM, New York, NY, USA, 450–462.

[58] K. T. Malladi, F. A. Nothaft, K. Periyathambi, B. C. Lee, C. Kozyrakis, and M. Horowitz. 2012. Towards energy-proportional datacenter memory with mobile DRAM. In *2012 39th Annual International Symposium on Computer Architecture (ISCA)*. 37–48. https://doi.org/10.1109/ISCA.2012.6237004

[59] Josiah McClurg, Raghuraman Mudumbai, and Joseph Hall. 2016. Fast demand response with datacenter loads. In *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, Minneapolis, MN, USA, 1–5. https://doi.org/10.1109/ISGT.2016.7781219

[60] Kendall Mongird, Vilayanur V. Viswanathan, Patrick J. Balducci, Md. Jan E. Alam, Vanshika Fotedar, V. S. Koritarov, and Boualem Hadjerioua. 2019. *Energy Storage Technology and Cost Characterization Report*. Technical Report. Pacific Northwest National Lab. (PNNL), Richland, WA (United States).

[61] Rajiv Nishtala, Vinicius Petrucci, Paul Carpenter, and Magnus Själander. 2020. Twig: Multiagent task management for colocated latency-critical cloud services. In *Proceedings of the International Symposium High-Performance Computer Architecture*.

[62] Amy Ousterhout, Joshua Fried, Jonathan Behrens, Adam Belay, and Hari Balakrishnan. 2019. Shenango: Achieving high CPU efficiency for latency-sensitive datacenter workloads. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI'19)*. USENIX Association, Boston, MA, 361–378. https://www.usenix.org/conference/nsdi19/presentation/ousterhout.

[63] Ali Pahlevan, Marina Zapater, Ayse K. Coskun, and David Atienza. 2020. ECOGreen: Electricity cost optimization for green datacenters in emerging power markets. *IEEE Transactions on Sustainable Computing* 6, 2 (2020), 289–305.

[64] T. Palit and M. Ferdman. 2016. Demystifying cloud benchmarking. In *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 122–132. https://doi.org/10.1109/ISPASS.2016.7482080

[65] Tirthak Patel and Devesh Tiwari. 2020. CLITE: Systematic and efficient co-location of multiple latency-critical and throughput-oriented workloads. In *Proceedings of the International Symposium High-Performance Computer Architecture*.

[66] PJM. 2022. Data Miner 2. https://dataminer2.pjm.com/list.

[67] PJM. 2022. PJM Manual 12: Balancing Operations. https://www.pjm.com/-/media/documents/manuals/m12.ashx.

[68] PJM. 2022. Real-Time Hourly LMPs. https://dataminer2.pjm.com/feed/rt_hrl_lmps/definition.

[69] George Prekas, Mia Primorac, Adam Belay, Christos Kozyrakis, and Edouard Bugnion. 2015. Energy proportionality and workload consolidation for latency-critical applications. In *Proceedings of the Sixth ACM Symposium on Cloud Computing*. 342–355.

[70] Kiran Ranganath, Joshua D. Suetterlein, Joseph B. Manzano, Shuaiwen Leon Song, and Daniel Wong. 2021. MAPA: Multi-accelerator pattern allocation policy for multi-tenant GPU servers. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–14.

[71] P. Ranganathan, P. Leech, D. Irwin, and J. Chase. 2006. Ensemble-level power management for dense blade servers. In *33rd International Symposium on Computer Architecture (ISCA'06)*. 66–77. https://doi.org/10.1109/ISCA.2006.20

[72] Francisco Romero and Christina Delimitrou. 2018. Mage: Online and interference-aware scheduling for multi-scale heterogeneous systems. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques*.

[73] Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, et al. 2020. Data center power oversubscription with a medium voltage power plane and priority-aware capping. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 497–511.

[74] Rasool Sharifi and Zainalabedin Navabi. 2017. Online profiling for cluster-specific variable rate refreshing in high-density DRAM systems. In *2017 22nd IEEE European Test Symposium (ETS)*.

[75] Navin Sharma, Sean Barker, David Irwin, and Prashant Shenoy. 2011. Blink: Managing server clusters on intermittent power. In *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems*. 185–198.

[76] Yuanyuan Shi, Bolun Xu, Baosen Zhang, and Di Wang. 2016. Leveraging energy storage to optimize data center electricity cost in emerging power markets. In *Proceedings of the Seventh International Conference on Future Energy Systems (e-Energy'16)*. ACM, New York, NY, USA, 18:1–18:13. https://doi.org/10.1145/2934328.2934346

[77] Chandrasekar Subramanian, Arunchandar Vasan, and Anand Sivasubramaniam. 2010. Reducing data center power with server consolidation: Approximation and evaluation. In *2010 International Conference on High Performance Computing*. 1–10. https://doi.org/10.1109/HIPC.2010.5713161

[78] Niraj Tolia, Zhikui Wang, Manish Marwah, Cullen Bash, Parthasarathy Ranganathan, and Xiaoyun Zhu. 2008. Delivering energy proportionality with non energy-proportional systems-optimizing the ensemble. *HotPower* 8 (2008), 2–2.

[79] Bhuvan Urgaonkar, Prashant Shenoy, Abhishek Chandra, and Pawan Goyal. 2005. Dynamic provisioning of multi-tier internet applications. In *Second International Conference on Autonomic Computing (ICAC'05)*. IEEE, 217–228.

[80] Balajee Vamanan, Hamza Bin Sohail, Jahangir Hasan, and T. N. Vijaykumar. 2015. TimeTrader: Exploiting latency tail to save datacenter energy for online search. In *Proceedings of the 48th International Symposium on Microarchitecture*.

[81] G. M. Wamba, Y. Li, A. Orgerie, N. Beldiceanu, and J. Menaud. 2017. Cloud workload prediction and generation models. In *2017 29th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. 89–96. https://doi.org/10.1109/SBAC-PAD.2017.19

[82] Di Wang, Chuangang Ren, Anand Sivasubramaniam, Bhuvan Urgaonkar, and Hosam Fathy. 2012. Energy storage in datacenters: What, where, and how much?. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'12)*. ACM, New York, NY, USA, 187–198. https://doi.org/10.1145/2254756.2254780

[83] Wei Wang, Amirali Abdolrashidi, Nanpeng Yu, and Daniel Wong. 2019. Frequency regulation service provision in data center with computational flexibility. *Applied Energy* 251 (2019), 113304.

[84] Xiaorui Wang, Ming Chen, Charles Lefurgy, and Tom W. Keller. 2011. SHIP: A scalable hierarchical power control architecture for large-scale data centers–supplementary file. (2011).

[85] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad. 2014. Opportunities and challenges for data center demand response. In *International Green Computing Conference*. 1–10. https://doi.org/10.1109/IGCC.2014.7039172

[86] Daniel Wong. 2016. Peak efficiency aware scheduling for highly energy proportional servers. In *International Symposium on Computer Architecture*.

[87] D. Wong and M. Annavaram. 2012. KnightShift: Scaling the energy proportionality wall through server-level heterogeneity. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*. 119–130. https://doi.org/10.1109/MICRO.2012.20

[88] Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. 2016. Dynamo: Facebook's data center-wide power management system. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 469–480.

[89] S. Wu, Y. Chen, X. Wang, H. Jin, F. Liu, H. Chen, and C. Yan. 2018. Precise power capping for latency-sensitive applications in datacenter. *IEEE Transactions on Sustainable Computing* (2018), 1–1. https://doi.org/10.1109/TSUSC.2018.2881893

[90] Hailong Yang, Alex Breslow, Jason Mars, and Lingjia Tang. 2013. Bubble-Flux: Precise online QoS management for increased utilization in warehouse scale computers. In *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA'13)*.

[91] Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2016. Elfen scheduling: Fine-grain principled borrowing from latency-critical workloads using simultaneous multithreading. In *2016 USENIX Annual Technical Conference (USENIX ATC'16)*.

[92] Chaojie Zhang, Alok Gautam Kumbhare, Ioannis Manousakis, Deli Zhang, Pulkit A. Misra, Rod Assis, Kyle Woolcock, Nithish Mahalingam, Brijesh Warrier, David Gauthier, et al. 2021. Flex: High-availability datacenters with zero reserved power. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 319–332.

[93] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen, and P. Li. 2018. An efficient deep learning model to predict cloud workload for industry informatics. *IEEE Transactions on Industrial Informatics* 14, 7 (July 2018), 3170–3178. https://doi.org/10.1109/TII.2018.2808910

[94] Y. Zhang, M. A. Laurenzano, J. Mars, and L. Tang. 2014. SMiTe: Precise QoS prediction on real-system SMT processors to improve utilization in warehouse scale computers. In *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'47)*. IEEE Computer Society, Washington, DC, USA, 406–418.

[95] Yijia Zhang, Ioannis Ch. Paschalidis, and Ayse K. Coskun. 2019. Data center participation in demand response programs with quality-of-service guarantees. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. 285–302.

[96] Yijia Zhang, Daniel C. Wilson, Ioannis Ch. Paschalidis, and Ayse K. Coskun. 2021. A data center demand response policy for real-world workload scenarios in HPC. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 282–287.

Received April 2021; revised February 2022; accepted March 2022