Applied Energy 251 (2019) 113304

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Frequency regulation service provision in data center with computational flexibility

Wei Wang*, Amirali Abdolrashidi, Nanpeng Yu*, Daniel Wong

Electrical and Computer Engineering, University of California, Riverside, Riverside, CA 92521, USA

HIGHLIGHTS

- A framework for data center to bid into electricity market and follow regulation signal in real-time.
- A risk-constrained bidding strategy is developed to determine the optimal energy and frequency regulation bids.
- Dummy load is introduced to increase the amount of regulation service provision of data center.
- Bi-linear server power consumption model and rule-based control enable data center to follow regulation signal accurately.
- The proposed regulation service provision framework reduces electricity bill by 12%.

ARTICLE INFO

Keywords:
Data center
Deep sleep state
Electricity market
Frequency regulation

ABSTRACT

The rapid adoption of cloud storage and computing services led to unprecedented growth of data centers in the world. As bulk energy consumers, large-scale data centers in the U.S. rack up billions in electricity costs annually. Fortunately, the operational flexibility of data centers can be leveraged to provide valuable frequency regulation services in smart grids to mitigate the indeterminacy of the renewable generation resources. Specifically, this paper aims to leverage computational flexibility provided by servers, such as dynamic voltage frequency scaling and dummy loads. This paper develops a comprehensive framework for data center's frequency regulation service provision in both hour-ahead market and real-time operations. A risk constrained hour-ahead bidding strategy along with a real-time data center power consumption control algorithm are developed to minimize electricity bills and the total response time of the requests. The introduction of dummy load, realistic bi-linear server power consumption model, and probabilistic forecast of electricity and frequency regulation service prices enable the data center to accurately follow frequency regulation signals, while reducing the financial risks associated with electricity market participation. The simulation results show that the proposed frequency regulation provision framework results not only in significant cost reduction for data centers, but also limits degradation in quality of service. Meanwhile, the stability and reliability of a power grid will be improved by the frequency regulation service provision.

1. Introduction

The emergence of cloud computing services drove the tremendous growth of data centers in the past ten years. Data centers have become a significant segment of the U.S. energy consumption. According to a recent U.S. data center energy usage report [1], around 70 TW h of electricity was consumed by data centers in 2014 and the annual shipment of data center servers is expected to grow 3% annually through 2020. The operational cost is a major component of the total cost of ownership of a data center. The electricity cost is about 30–50% of the total operational cost [2]. Therefore, it is imperative to improve the operations of the data centers to lower the electricity bill.

The power management of data centers has been widely studied from different aspects. Dynamic voltage and frequency scaling (DVFS) [3–5], virtual machine migration and auto-scaling [6,7], and geometrical load balancing [8] techniques have been widely explored to lower power consumption in data center servers. To lower power in cooling systems, both emission-efficient and energy efficient economizers are developed in [9–12].

Renewable energy is the fastest-growing fuel source of power systems in the past ten years [13,14]. The intermittency of renewable generation outputs poses a new set of operational and planning

* Corresponding authors.
E-mail addresses: wwang031@ucr.edu (W. Wang), aabdo001@ucr.edu (A. Abdolrashidi), nyu@ece.ucr.edu (N. Yu), dwong@ece.ucr.edu (D. Wong).

https://doi.org/10.1016/j.apenergy.2019.05.107
Received 23 January 2019; Received in revised form 2 May 2019; Accepted 7 May 2019

Nomenclature

$\alpha_1^j, \alpha_2^j, \alpha_3^j, \alpha_3^j$	coefficients of fitted power curve in j-th range
δ_{risk}	threshold of risk constraint
μ_1, μ_2, μ_3	Lagrange multipliers for capacity bidding constraints
$\overline{C}_{dif}(t)$	expectation of price difference at time t
B_{cap}	bidding capacity for frequency regulation service
$B_{cap}(t)$	bidding capacity for frequency regulation service at time t
$C_{dif}(t)$	price difference between effective electricity price and regulation service price at time t
C_{dif}^i	price difference of data sample i
$C_{efe}(t)$	effective energy price considering cooling cost at time t
$C_e(t)$	electricity price at time t
$C_{reg}(t)$	regulation service price at time t
$cap_{max}(f)$	maximum computing capacity of a single server with frequency f
$dpr_i(t)$	ratio of dynamic power consumption to request per second rate of server i at time t
$f_i(t)$	frequency of server i at time t
f_{max}	maximum frequency of server
N	total number of servers in the data center
N_d	total number of training data samples
P_0	deep sleep state idle power consumption of a single server when utilization rate is 0
$P_i(t)$	power consumption of server i at time t
P_{base}	power consumption base submitted to the hour-ahead market
$P_{DC}(t)$	total power consumption of data center at time t
$P_{DC}^{base}(t)$	power consumption base of data center at hour t
P_{DC}^{max}	maximum power consumption of the data center
$P_{DC}^{min}(t)$	minimum power consumption of the data center at time t
$P_{set}(t)$	power set point for frequency regulation at time t
$P_{uni}(t)$	total power consumption of data center with uniform routing at time t
$r(t)$	total number of requests arrive at the data center per second at time t
$r_i(t)$	the number of requests routed to server i per second at time t
$RegD(t)$	fast regulation signal at time t
$rt_i(t)$	request response time of server i at time t
$rt_{SLA}(t)$	response time limit under service level agreement at time t
$score(t)$	performance score at time t
$u_i(t)$	total utilization rate of server i at time t
$u_{uni}(t)$	utilization rate of each server by uniformly routing the requests at time t
$ud_i(t)$	utilization rate of server i by dummy load at time t
$ur_i(t)$	utilization rate of server i by request at time t

challenges to power system operators. In particular, there is an increasing need for high quality frequency regulation services to balance the supply and demand of electricity in real-time and mitigate the uncertainties in renewable generation outputs [15]. A major challenge of automatic generation control (AGC) in fossil-fueled power plants is that they are not well suited to follow the AGC set points on a second-by-second basis with very high accuracy.

The use of data centers to provide frequency regulation service has attracted a great deal of interest recently. Data centers need to participate in two electricity market processes to provide frequency regulation services: the hour-ahead (HA) market bidding process and the real-time operations. The existing literature can be divided into two groups based on which market/operation process was considered.

In the first group of literature, the profit maximization problem of the data center is formulated to determine the optimal bidding strategy for energy and frequency regulation services. In [16], an optimization-based profit maximization problem for data centers with quality of service (QoS) constraint is formulated. The service rate is controlled to offer load reduction as an ancillary service. In [17], the problem of leveraging energy storage systems in data centers to provide frequency regulation service and peak shaving service is studied. However, the profit maximization problem formulated in existing literature did not take the uncertainty of energy, frequency regulation service prices, and data center requests arrival rates into consideration. In addition, the financial risks associated with participating in the electricity market are not modeled.

In the second group of literature, the real-time frequency regulation signal following problem of data centers is investigated. In [18,19], the DVFS and CPU resource limit techniques are adopted to adjust the power consumption. Different real-time control policies, such as efficiency-first and priority-first policies, are proposed to follow the frequency regulation signal. Various power states of the servers are considered in [20,21] to offer additional flexibility for power control, including active, idle, slow-to-wakeup sleep state, and shut-down power states. In [22], a stochastic dynamic programming problem is formulated to find the optimal policy for the frequency regulation service provision while reducing the quality of service degradation. In [23,24], battery storage systems are leveraged to provide frequency regulation services. Peak demand reduction of the data center is also considered in [23]. In [25,26], both CPU frequency and the charging schedule of electric vehicles are controlled to follow the frequency regulation signals in real-time. Our proposed approach does not rely on batteries in data centers or electric vehicles to follow frequency regulation. Instead, we rely solely on computational flexibility provided by computational resources, such as DVFS and server load. A main novel contribution of this work is to explore the potential of providing additional frequency regulation services by introducing dummy computing load. Therefore, our frequency regulation policy can be implemented as simple software runtimes to “reshape” the power consumption of the server. This approach is complementary to battery-based approaches, and does not result in battery lifetime issues common in battery-based frequency regulation approaches.

Another limitation of prior literature is that most of the existing literature adopted a simplified linear power consumption model for servers.In our work, we show that simple linear power models do not allow data centers to accurately follow real-time frequency regulation signals, potentially resulting in increased electricity bills when providing frequency regulation services.

In this paper, we propose a comprehensive framework for the frequency regulation provision by data centers, which covers the hour-ahead market bidding and the real-time signal following problems. A piece-wise bi-linear energy consumption model of the data center servers with default deep sleep state policy is first derived based on empirical measurements from real-world tests on servers. Dummy computing loads are introduced to control server power consumption in addition to the traditional DVFS technique. A neural network-based probabilistic model of energy and frequency regulation service prices is developed and embedded into the risk constrained optimization problem to determine the optimal energy and frequency regulation service bids for the data center in the hour-ahead market. For real-time operations, a rule-based data center power consumption control algorithm is developed, which not only enables frequency regulation signal following with high accuracy but also reduces the total response time of the requests.

The unique contributions of this paper are listed as follows:

- This paper provides a comprehensive framework for a data center to bid into the hour-ahead electricity market and follow frequency

regulation signal in real-time operations.

- A risk-constrained hour-ahead bidding strategy considering uncertainties of energy and frequency regulation service prices is developed to determine the optimal energy and frequency regulation bids by data centers.
- The dummy computing load is introduced for the first time to increase the amount of frequency regulation service provision of the data center in addition to the DVFS technique.
- A realistic bi-linear server power consumption model and rule-based data center power consumption control algorithm not only enable accurate frequency regulation signal following but also limit degradation in QoS.
- The theoretical derivation and simulation results point out that the profitability of frequency regulation service provision by data center depends on accurate prediction of the price difference between frequency regulation service and energy.
- The simulation results show that for a period of 3 months, the proposed frequency regulation service provision framework reduces the electric bill by $21,590 (8.1%) for a data center with 100,000 servers compared to the power minimization strategy.

The remainder of this paper is organized as follows: Section 2 lays out the overall framework for the data center to participate in the electricity market to provide regulation services. Section 3 develops the energy consumption model of realistic servers with real-world power measurement data. Section 4 and 5 formulate the optimization problem of hour-ahead capacity bidding and real-time signal following respectively. A risk-limited bidding strategy and a rule-based signal-following algorithm are proposed. The numerical simulations with the Wikipedia requests for workload trace and price data from PJM market are performed in Section 6. Finally, the paper is concluded in Section 7.

2. Overall framework

The overall framework of the frequency regulation service provision by a data center is depicted in Fig. 1. The overall framework involves interactions between a transmission system operator (TSO) and a data center (DC) in two electricity market processes: hour-ahead market and real-time operations. The details of the frequency regulation service provision framework is described in the next three subsections. The proposed framework is applicable to different electricity markets. The specific implementation of the bidding strategy can be easily adjusted to different market rules.

2.1. Data center's participation in electricity market

As shown in Fig. 1, in order to provide frequency regulation services, the data center is required to participate in two electricity market processes: HA market and real-time operations.

In the HA market, the data center will first predict the prices for energy and frequency regulation services, and the workload of the data center for the next operating hour. The data center will then determine the optimal bidding capacity for energy and frequency regulation services which maximize its expected net benefits subject to certain risk limits. After the HA market is cleared by the transmission system operator, the data center will receive the hour-ahead energy schedule, the award for frequency regulation service, and the cleared prices for energy and frequency regulation.

In real-time operations, the data center receives the frequency regulation signals and automatic generation control (AGC) set points from the transmission system operator every 2 s. The frequency regulation signals range from -1 to 1. The signals are negative (positive) when the system requests frequency regulation down (up) services. The AGC set points specify the amount of load the data center should consume. The AGC set points are equal to the summation of the HA market energy schedule plus the product of the frequency regulation signals and frequency regulation service awards. Upon receiving the AGC set points, the data center adjusts its energy consumption to follow the set points. Data centers can accurately follow the AGC set points by dynamically routing arriving requests to various servers, changing the operating frequency of CPUs and inserting dummy loads at the server level.

The physical and contractual constraints of the data center need to be taken into consideration when participating in the electricity market. First, the bidding capacity for energy P_{base} and frequency regulation service B_{cap} should be determined in such a way that the maximum and minimum power consumption limits P_{max} and P_{min} of the data center will not be violated. If the submitted bids are accepted, then in real-time operations the AGC set points for the data center ranges from $P_{base} - B_{cap}$ to $P_{base} + B_{cap}$. The data center needs to make sure $P_{base} + B_{cap} \leqslant P_{max}$ and $P_{base} - B_{cap} \geqslant P_{min}$. Second, as a cloud computing service provider, the data center also needs to satisfy the service level agreement (SLA) and maintain the QoS. Hence, the control of request routing, CPU frequency, and dummy loads are limited by the SLA requirements.

Finally, note that in electricity markets such as Pennsylvania-New Jersey-Maryland Interconnection (PJM), there are two types of frequency regulation services, *RegA* and *RegD*. The real-time regulation signal of *RegD* service is much more volatile than that of *RegA* service, and the price of *RegD* service is higher than that of *RegA* service. The data center is capable of controlling its server energy usage in real-time to follow the volatile *RegD* service signals. Hence, it is suitable for the data center to provide such premium frequency regulation services and receive higher compensation from the electricity market. Fig. 2 shows an example of daily prices for frequency regulation services and energy in the PJM market. For about 32% of hours in year 2017 and 2018, the frequency regulation price of RegD service is higher than the energy price in PJM market. As the penetration level of renewable energy increases, the demand for frequency regulation service will rise as well. This will further increase the percentage of hours where the frequency regulation service price is higher than the energy price.

2.2. Transmission system operator

In the hour-ahead market process, the transmission system operator first receives both energy and frequency regulation service bids from generators and data centers. It then clears the hour-ahead market to determine the hour-ahead energy schedule and prices for energy and frequency regulation services. The objective is to minimize the total energy and frequency regulation service costs while satisfying the electric loads [27]. The market clearing results will be sent to data centers and other market participants. In the real-time operations, the transmission system operator will first measure area control error and compute the frequency regulation signals of the system aiming to reduce the area control error to zero in a distributed fashion[28]. The individual generator and data center's AGC set points will be calculated based on the frequency regulation signal, hour-ahead energy schedule, and frequency regulation service awards. The updated AGC set points

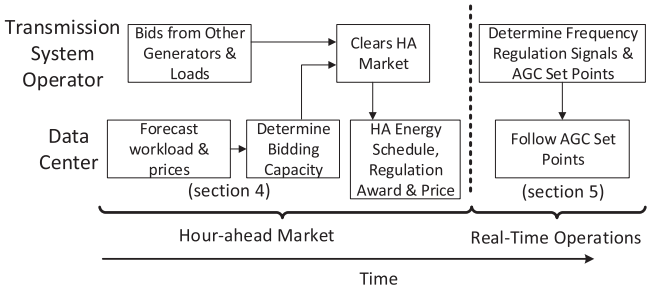

Fig. 1. Overall framework of frequency regulation service provision by data center.

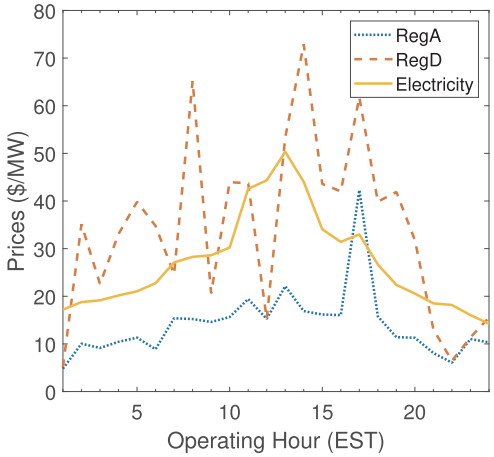

Fig. 2. Prices for frequency regulation services and energy in PJM market.

will be sent to the generators and data centers every 2 s.

2.3. Performance-based compensation

The final compensation for providing frequency regulation services depends on both the frequency regulation service award amount and the real-time AGC set point signal following performance. The signal following performance is quantified by the performance score in PJM market [29]. It consists of three components: accuracy, delay, and precision.

The accuracy score is the correlation between the AGC set point signals and data center's response. It is calculated over a five-minute period with 10-s granularity. The calculation is performed repeatedly with 10-s delays propagated over five minutes, where the best score is used. The delay score is based on the time delay between the control signal and the point of the highest correlation. The delay score will be 100% if the best correlation is at 0 or 10-s delay. It decreases as the delay time increases until the 5-min mark. The precision score is calculated based the instantaneous error between the control signal and the regulating unit's response. The final performance score is the average of the three components.

3. Energy model of server

To build an energy model of a server, we empirically profiled a server running CentOS 7 with 32 Intel Xeon E5 cores across different working frequencies between 1.2 GHz and 2.1 GHz (maximum frequency without turbo-boosting) with the default C6 sleep policy. We focus the energy model on processors because processors are the largest consumers of power in data centers and have the largest dynamic range [30]. In addition, it has been widely observed that processor power consumption can be used as a proxy for whole-server power [30–32].

As a workload, we used the Web Search benchmark from Cloudsuite [33] with a ramp time and steady state time of 30 s and 250 s respectively. The measurements taken from the tests include the average power consumption and percentage of idle time. To find the maximum computing load a server can handle, we gradually increase the number of clients in the benchmark at every frequency until it fails to satisfy QoS and pick the greatest value. The idleness and power measurements are performed using *powertop* and *rapl-read*[34] respectively, while the frequency is scaled using *cpufreq* drivers.

The percentage of C6 sleep time and CPU power consumption of a single server with different request rates (per second) are depicted in Fig. 3. As shown in the figure, the default C6 sleep time percentage decreases with the increase of request per second (RPS) almost linearly for each frequency. The C6 sleep time of a server running at maximum frequency f_{max}, i.e. 2.1 GHz, reaches zero at 1230 RPS, which is deemed as the maximum capacity of the server $cap_{max}(f_{max})$. Fig. 3 also shows the maximum capacity of the server under various frequencies $cap_{max}(f)$ increases approximately linearly with the frequency of the server CPU.

Hence, we can estimate the maximum capacity of the server under frequency f with:

$$cap_{max}(f) = \frac{f}{f_{max}} cap_{max}(f_{max}) \tag{1}$$

In Fig. 3, the circles on the horizontal axis represent the estimated maximum capacity under different frequencies by scaling cap_{max} with the above linear equation. The squares on the horizontal axis of the figure are the approximated maximum capacity by linear extrapolation with the last 4 points of each curve. The short distances between the circles and squares show that the linear approximation for the maximum server capacity is fairly accurate.

The utilization rate of the server i at time t is defined as the ratio of the number of requests per second to the maximum server capacity under a particular frequency:

$$ur_i(t) = \frac{r_i(t)}{cap_{max}(f_i(t))} \tag{2}$$

The energy consumption model with the default sleep policy is the baseline considered in this paper. At the server level, the energy consumption can be controlled by adjusting the CPU frequency and introducing the dummy computing load. The dummy load can be trivially injected by running a process that stresses the CPU with mainly compute instructions, limiting performance interference with other processes running in the server. It increases the equivalent utilization rate and decreases the sleep time percentage. The total equivalent utilization rate with the dummy load is

$$u_i(t) = ur_i(t) + ud_i(t) \tag{3}$$

The power consumption of a single server with the default sleep policy at different frequencies and utilization rates are depicted by the markers in Fig. 4. The relationship between the power consumption and utilization rate of the servers can be described by a piece-wise bi-linear function. Note that the slope of the power curve segment where the utilization rate is between 0 and 0.1 is larger than that of the segment where the utilization rate is between 0.1 and 1. It can also be observed that for a fixed utilization rate, the power consumption increases faster when the frequency increases from 2.0 GHz to 2.1 GHz than when the frequency increases from 1.2 GHz to 2.0 GHz.

Therefore, the server power can be approximated as a piece-wise bi-linear function of frequency and utilization rate as:

$$P_i(t) = \alpha_1^j f_i(t) u_i(t) + \alpha_2^j u_i(t) + \alpha_3^j f_i(t) + \alpha_4^j \tag{4}$$

where the four different ranges of utilization rate and frequency are defined as follows:

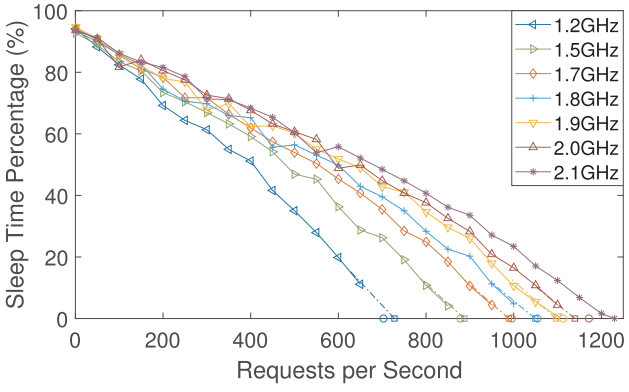

Fig. 3. C6 sleep time versus request per second rate.

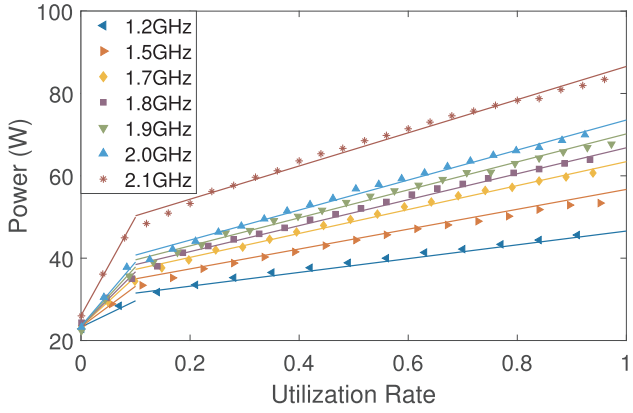

Fig. 4. Fitted power consumption curves with default sleep policy.

$$j = \begin{cases} 1, & 0 < = u_i(t) < = 0.1,\ 1.2 < = f_i(t) < = 2.0 \\ 2, & 0 < = u_i(t) < = 0.1,\ 2.0 < = f_i(t) < = 2.1 \\ 3, & 0.1 < = u_i(t) < = 1,\ 1.2 < = f_i(t) < = 2.0 \\ 4, & 0.1 < = u_i(t) < = 1,\ 2.0 < = f_i(t) < = 2.1 \end{cases} \tag{5}$$

The power consumption curves fitted with least square regression according to Eq. (4) are depicted in Fig. 4. As shown in the figure, the piece-wise bi-linear function is well suited to model the server power consumption.

This piece-wise bi-linear model is typical of modern processors. Leakage energy dominates at low utilization levels (below 10%), and idle power management policies, such as sleep states and circuit-level power gating, have a profound effect. Furthermore, at 2.1 GHz we observed that power increases at a faster rate. This is also typical of modern processors that utilize frequency boosting techniques, such as Intel TurboBoost or AMD Turbo Core [35]. These policies aim to maximize performance and leverage the processor's thermal headroom. Therefore, operating at these frequencies sacrifices energy efficiency for performance, thus the increase in power consumption rate.

4. Hour-ahead market frequency regulation and energy bidding strategy for data center

4.1. Problem formulation

In the hour-ahead market, the objective of the data center is to determine the optimal bidding strategy which maximizes the expected net earnings for each hour *t*. The net earnings of the data center can be calculated as the difference between the revenue received from frequency regulation service provision and the total electricity cost as shown in Eq. (6):

$$\max E\left[C_{reg}(t)score(t)B_{cap}(t) - (1+\eta_{cool})C_e(t)P_{DC}^{base}(t)\right] \tag{6}$$

where $C_e(t)$ and $C_{reg}(t)$ are electricity price and frequency regulation price for hour *t*. The cooling cost coefficient, i.e., ratio of cooling power over server load is denoted by $\eta_{cool} \in R^+$. The decision variables are the bidding capacity for frequency regulation service $B_{cap}(t)$ and energy consumption $P_{DC}^{base}(t)$. Note that the data center is assumed to be a price-taker in the electricity market participation process.

The constraints for the net earnings maximization problem are as follows:

$$P_{DC}^{min}(t) \leqslant P_{DC}^{base}(t) - B_{cap}(t) \tag{7}$$

$$B_{cap}(t) + P_{DC}^{base}(t) \leqslant P_{DC}^{max} \tag{8}$$

$$f_{risk}(C_{reg}(t),\ C_{efe}(t),\ r(t),\ B_{cap}(t),\ P_{DC}^{base}(t)) \leqslant \delta_{risk} \tag{9}$$

where $C_{efe}(t) = (1+\eta_{cool})C_e(t)$ is the effective energy price considering cooling cost. Eqs. (7) and (8) represent the upper and lower bidding capacity constraints. Eq. (9) represents the risk limit constraint for the data center's bidding strategy, where $r(t)$ represents the average request arrival rate during hour *t*.

If the data center does not provide frequency regulation service, then its optimal bidding strategy aims at minimizing energy cost. Hence, the risk of the joint energy and frequency regulation service bidding strategy can be defined as the expectation of the bidding strategy loss compared to the power consumption minimization scenario. Note that losses arise in cases when the revenue received from frequency regulation service provision is less than the increased energy cost:

$$\begin{aligned} &f_{risk}(C_{reg}(t),\ C_{efe}(t),\ r(t),\ B_{cap}(t),\ P_{DC}^{base}(t)) \\ &= -\iint_V Pr(C_{reg}(t),\ C_{efe}(t))\{[C_{reg}(t)score(t)B_{cap}(t) - \\ &C_{efe}(t)P_{DC}^{base}(t)] - (-C_{efe}(t)P_{DC}^{min}(t))\}dC_{reg}(t)dC_{efe}(t) \\ &= \iint_V Pr(C_{reg}(t),\ C_{efe}(t))\{C_{efe}(t)[P_{DC}^{base}(t) - P_{DC}^{min}(t)] \\ &- C_{reg}(t)score(t)B_{cap}(t)\}dC_{reg}(t)dC_{efe}(t) \end{aligned} \tag{10}$$

where V is defined as follows:

$$V = \{C_{reg}(t),\ C_{efe}(t) | C_{efe}(t)[P_{DC}^{base}(t) - P_{DC}^{min}(t)] > C_{reg}(t)score(t)B_{cap}(t)\} \tag{11}$$

$Pr(C_{reg}(t),\ C_{efe}(t))$ denotes the joint probability distribution of frequency regulation service price $C_{reg}(t)$ and effective energy price $C_{efe}(t)$.

In Eq. (8), P_{DC}^{max} denotes the maximum power consumption of the data center, which can be calculated by summing up individual servers' power consumption at full utilization rate and maximum CPU frequency as $P_{DC}^{max} = \sum_{i=1}^{N} P_i(f = 2.1,\ u = 1.0)$.

In Eqs. (7) and (10), the minimum power consumption of data center P_{DC}^{min}, can be found by solving the following optimization problem with $r_i(t)$, $u_i(t)$, and $f_i(t)$ as decision variables:

$$P_{DC}^{min}(t) = \min \sum_{i=1}^{N} P_i(t) \tag{12}$$

s.t.

$$\sum_{i=1}^{N} r_i(t) = E\left[r(t)\right] \tag{13}$$

$$P_i(t) = \alpha_1^j f_i(t)u_i(t) + \alpha_2^j u_i(t) + \alpha_3^j f_i(t) + \alpha_4^j \tag{14}$$

$$u_i(t) = ur_i(t) \leqslant 100\% \tag{15}$$

$$ur_i(t) = \frac{r_i(t)}{cap_{max}(f_i(t))} \tag{16}$$

$$rt_i(t) = f_{rt}(f_i(t),\ u_i(t)) \leqslant rt_{SLA} \tag{17}$$

The objective function (12) aims at minimizing the summation of the power consumption of each server. Eqs. (13)–(17) represent the operation constraints of the data center. We assume a homogeneous computing environment in the data center. Hence, Eq. (13) ensures that the summation of requests routed to each server should be equal to the total requests received by the data center. Eq. (14) represents the power consumption model of each server where only the CPU power is considered. As shown in Section 3, the power consumption of a server with the default sleep state policy is a piece-wise bi-linear function of CPU frequency and utilization rate. Eq. (15) enforces the upper limit of the CPU utilization rate. Note that in the power consumption minimization problem, the dummy computing load must be zero. Hence $u_i(t) = ur_i(t)$. Eq. (17) represents the service level agreement constraint which sets upper limits on the response time of the 90th percentile of the requests. As shown in Fig. 5, the response time is a function of utilization rate. Hence, we can set an equivalent upper bound on the utilization rate $u_i(t)$. For example, the utilization rate limit of $rt_{SLA} = 115$ ms at $f = 2.1$ GHz is about 0.8. Define $u_{max}(f)$ as the corresponding utilization

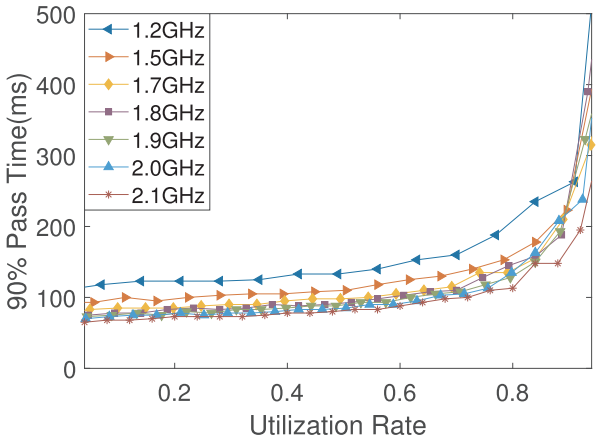

Fig. 5. 90% pass time versus utilization rate.

limit of the SLA. Although the explicit functional form of the utilization rate limit is not modeled here, the estimation for the minimum power consumption P_{DC}^{min} can still be performed as shown in the Appendix C.

4.2. Problem reformulation

The energy price C_e is rarely negative in practice. In PJM market, there are only 23 hours in total with negative energy prices during year 2017 and 2018. When the energy price becomes negative, the net earnings of a data center when providing both frequency regulation and energy services are greater than when the data center tries to minimize power consumption. Hence, we only consider the cases when $C_{efe} > 0$.

The Lagrange function of the optimization problem (6)–(9) is

$$\begin{aligned}\mathscr{L} &= E[C_{efe}(t)]P_{DC}^{base}(t) - E[C_{reg}(t)score(t)]B_{cap}(t) + \mu_1 \\ &(P_{DC}^{min}(t) + B_{cap}(t) - P_{DC}^{base}((t)) + \mu_2(B_{cap}(t) + P_{DC}^{base}((t) - P_{DC}^{max}(t)) \\ &+ \mu_3(f_{risk} - \delta)\end{aligned} \tag{18}$$

where $\mu_1 \geqslant 0$, $\mu_2 \geqslant 0$, and $\mu_3 \geqslant 0$ are the corresponding Lagrange multipliers. By taking partial derivative of the Lagrange function with respect to $P_{DC}^{base}(t)$, we obtain

$$\frac{\partial \mathscr{L}}{\partial P_{DC}^{base}(t)} = E\left[C_{efe}(t)\right] - \mu_1 + \mu_2 + \mu_3 \frac{\partial f_{risk}}{\partial P_{DC}^{base}(t)} \tag{19}$$

As shown in Appendix A, $(\partial f_{risk})/(\partial P_{DC}^{base}(t)) \geqslant 0$. Intuitively, the risk of providing frequency regulation service increases with $P_{DC}^{base}(t)$ because the electricity cost increases with $P_{DC}^{base}(t)$.

At the optimum point, (19) is equal to zero. Hence, we have

$$E\left[C_{efe}(t)\right] = \mu_1 - \mu_2 - \mu_3 \frac{\partial f_{risk}}{\partial P_{DC}^{base}(t)} > 0 \tag{20}$$

Therefore, at optimal solutions,

$$\mu_1 > \mu_2 + \mu_3 \frac{\partial f_{risk}}{\partial P_{DC}^{base}(t)} \geqslant 0 \tag{21}$$

Hence, the constraint (7) is binding at the optimum point, i.e.,

$$P_{DC}^{base}(t) = B_{cap}(t) + P_{DC}^{min}(t) \tag{22}$$

Therefore, the objective function (6) can be reformulated as:

$$\begin{aligned}&E[C_{reg}(t)score(t)B_{cap}(t) - C_{efe}(t)P_{DC}^{base}(t)] \\ &= E[C_{reg}(t)score(t) - C_{efe}(t)]B_{cap}(t) \\ &- E[C_{efe}(t)]P_{DC}^{min}(t)\end{aligned} \tag{23}$$

Let's define the price difference $C_{dif}(t)$ as $C_{dif}(t) = C_{reg}(t)score(t) - C_{efe}(t)$, and the expectation of $C_{dif}(t)$ as $\overline{C}_{dif}(t) = E[C_{dif}(t)]$. If the estimator of $r(t)$ is unbiased, then $P_{DC}^{min}(t)$ calculated based on $r(t)$ is also unbiased. Note that the second term on the right hand side of Eq. (23) is not a function of the decision variable $B_{cap}(t)$. Hence, maximizing the objective function (23) is equivalent to maximizing $\overline{C}_{dif}(t)B_{cap}(t)$, i.e., the extra benefits of providing frequency regulation services compared to the minimum power consumption strategy.

Similarly, by leveraging the equality (22), the risk limit constraint (9) can be simplified as

$$-\int_{-\infty}^{0} Pr\left(C_{dif}(t)\right)C_{dif}(t)B_{cap}(t)dC_{dif}(t) \leqslant \delta_{risk} \tag{24}$$

In summary, the optimization problem (6)–(9) can be reformulated as:

$$\max \overline{C}_{dif}(t)B_{cap}(t) \tag{25}$$

s.t.

$$-\int_{-\infty}^{0} Pr\left(C_{dif}(t)\right)C_{dif}(t)B_{cap}(t)dC_{dif}(t) \leqslant \delta_{risk} \tag{26}$$

$$0 \leqslant B_{cap}(t) \leqslant \frac{P_{DC}^{max}(t) - P_{DC}^{min}(t)}{2} \tag{27}$$

4.3. Solution methodology

In order to solve the optimization problem (25)–(27), the probability distribution of the price difference C_{dif} needs to be modeled and estimated first.

A feed-forward neural network can be trained to estimate the probability distribution of C_{dif} based on the observed price differences and input features X such as the historical prices, load and generation information. The conditional distribution of price difference given observed features and trained neural network parameters is assumed to be Gaussian.

$$Pr(C_{dif}|X, W) = \mathscr{N}(C_{dif}|y(X, W), \beta) \tag{28}$$

Where X denotes the input features, W denotes the weights of the neural network, and β denotes the variance of the Gaussian noise. $y(X, W)$ denotes the output of the neural network, which is the mean value of price difference variable which follows the Gaussian distribution.

Given a data set of N_d independent, identically distributed observations along with corresponding target values for price differences $\{(x_i, C_{dif}^i),\ i = 1, 2...,N_d\}$, we can construct the corresponding negative logarithm of the likelihood function:

$$\frac{1}{2\beta}\sum_{i=1}^{N_d}(y(x_i, W) - C_{dif}^i)^2 - \frac{N_d}{2}\ln\frac{1}{\beta} + \frac{N_d}{2}\ln\left(2\pi\right) \tag{29}$$

The weights of the neural network W can be obtained by maximizing the likelihood or minimizing the sum-of-square error function given by

$$\sum_{i=1}^{N_d}(y(X, W) - C_{dif}^i)^2 \tag{30}$$

Denote the W obtained by minimizing the sum-of-square error as W_{ml}. By making the partial derivative of Eq. (29) with respect to β equal to zero, β_{ml} can be obtained as

$$\beta_{ml} = \frac{\sum_{i=1}^{N_d}(y(x_i, W_{ml}) - C_{dif}^i)^2}{N_d} \tag{31}$$

After training the neural network and obtaining the network parameters W_{ml} and β_{ml}, we have $\overline{C}_{dif}(t) = y(x(t), W_{ml})$, and

$$Pr(C_{dif}(t)|x(t), W_{ml}) = \mathscr{N}(C_{dif}(t)|\overline{C}_{dif}(t), \beta_{ml}). \tag{32}$$

Therefore, the closed form solution of optimization problem (25)–(27)

is

When $\overline{C}_{dif}(t) > 0$,

$$B_{cap}(t) = min\left\{\frac{P_{DC}^{max}(t) - P_{DC}^{min}(t)}{2}, \frac{-\delta_{risk}}{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\beta_{ml}}} e^{-\frac{\left(C_{dif}(t) - \overline{C}_{dif}(t)\right)^2}{2\beta_{ml}}} C_{dif}(t) dC_{dif}(t)}\right\} \tag{33}$$

Otherwise $B_{cap}(t) = 0$. Therefore, the actual gain is influenced by the accuracy of the price difference forecast model as the $B_{cap}(t)$ is determined by $\overline{C}_{dif}$ and β_{ml}.

Note that to obtain $P_{DC}^{min}(t)$ in (33), the requests arrival rate needs to be modeled and estimated. In this paper, we adopt the auto-regressive integrated moving average (ARIMA) [36] model to approximate the time series of requests arrival rate.

5. Following real-time frequency regulation signal

5.1. Problem formulation

In the real-time frequency regulation signal following stage, the goal of the data center is to minimize the total response time of the requests while following the frequency regulation signals accurately. Therefore, the task of following real-time frequency regulation signal is equivalent to solving the following optimization problem:

$$\min \sum_{i=1}^{N} r_i(t) rt_i(t) \tag{34}$$

s.t.

$$P_{set}(t) - P_{DC}(t) = 0 \tag{35}$$

where $P_{set}(t) \triangleq P_{DC}^{base}(t) + B_{cap}(t) RegD(t)$ is the AGC set point sent to the data center by the transmission system operator. $P_{DC}^{base}(t)$ is the energy dispatch level of the data center and $B_{cap}(t) RegD(t)$ is the amount of frequency regulation service the data center is required to provide. Note that $RegD(t)$ is the frequency regulation signal which ranges from -1 to 1. $rt_i(t)$ is the average request response time of server i. The response time of each request is determined by the utilization rate of each server. At last, $P_{DC}(t) \triangleq \sum_{n=1}^{N} P_i(t)$, where $P_i(t)$ can be calculated by Eq. (14). The decision variables of the optimization problem include CPU frequency (discrete variable) and the dummy load (continuous variable) of each server.

5.2. Rule-based data center power consumption control algorithm

In order to make online adjustments of total power consumption as the frequency regulation signal is updated every 2 s, a rule-based control strategy is proposed.

The uniform server utilization rate at time t is defined by routing requests evenly to all servers operating with a CPU frequency of 2.1 GHz as

$$u_{uni}(t) = \frac{r(t)/N}{cap_{max}(f = 2.1)} \tag{36}$$

The total power consumption at time t by uniformly routing requests to all servers running at 2.1 GHz can be obtained as

$$P_{uni}(t) = \sum_{i}^{N} P_i\left(f = 2.1,\ u = u_{uni}(t)\right) \tag{37}$$

In order to accurately follow the frequency regulation signals while minimizing the total request response time, the operating strategy of the data center varies according to the AGC set point and total number of requests received by the data center as follows:

1. $P_{set}(t) \geqslant P_{uni}(t)$;
2. $P_{set}(t) < P_{uni}(t)$ and $r(t) \leqslant u_{max}(2.0) \times N \times cap_{max}(f = 2.0)$;
3. $P_{set}(t) < P_{uni}(t)$ and $r(t) > u_{max}(2.0) \times N \times cap_{max}(f = 2.0)$.

The operating strategy of the data center under each of the three cases will be presented in detail.

CASE 1: When $P_{set}(t) \geqslant P_{uni}(t)$, the minimum request response time can be achieved by uniformly routing requests to all the servers and adding dummy loads until the AGC set point is met. A proof for why the proposed data center operating strategy achieves minimum request response time while accurately following the frequency regulation signal is provided in the Appendix B.

To follow the frequency regulation signals accurately, dummy computing loads need to be added to increase the server utilization rate. The amount of dummy load $u_d(t)$ needed can be calculated as follows:

$$u_d(t) = u(t) - u_{uni}(t) \tag{38}$$

where $u(t)$ can be found by solving:

$$P_{set}(t) = \sum_{i=1}^{N} P_i\left(f = 2.1,\ u(t)\right) = N\left(2.1\alpha_1^j u(t) + \alpha_2^j u(t) + 2.1\alpha_3^j + \alpha_4^j\right) \tag{39}$$

Hence, the closed form solution of $u(t)$ is as follows:

$$u(t) = \frac{P_{set}(t) - N(2.1\alpha_3^j + \alpha_4^j)}{N(2.1\alpha_1^j + \alpha_2^j)} \tag{40}$$

Where

$$j = \begin{cases} 2, & P_{set}(t)/N < P_i(f = 2.1,\ u = 0.1) \\ 4, & P_{set}(t)/N \geqslant P_i(f = 2.1,\ u = 0.1) \end{cases} \tag{41}$$

CASE 2: When $P_{set}(t) < P_{uni}(t)$ and $r(t) \leqslant u_{max}(2.0) \times N \times cap_{max}(f = 2.0)$, the data center's operating strategy works as follows. Note that in this case the power consumption needs to be reduced from $P_{uni}(t)$ to $P_{set}(t)$ and the SLA can be satisfied with all servers running at 2.0 GHz. We will start from the baseline operating strategy where the requests are uniformly routed to all servers running at $f = 2.1$ GHz. Then we select n servers whose requests are packed to n' servers running at 2.0 GHz with the maximum utilization rate $u_{max}(2.0)$ which does not violate the SLA. The remaining $n - n'$ servers are kept in idle state. By carefully choosing n and n', the data center is capable of closely following the frequency regulation signals. Although the proposed data center operating strategy increased the response time for some of requests compared to the uniform routing benchmark, it minimizes the number of requests with increased response time as shown in Appendix D.

The data center control parameters n and n' can be calculated online as follows.

The total energy consumption of the data center in CASE 2 include energy consumption from n' servers with packed requests, $n - n'$ servers in idle, and $N - n$ servers operating under the uniform routing and the maximum frequency.

$$P_{DC}(t) = \sum_{i=1}^{n'} P_i\left(f = 2.0,\ u = u_{max}(2.0)\right) + \sum_{i=n'+1}^{n} P_0 + \sum_{i=n+1}^{N} P_i\left(f = 2.1,\ u = u_{uni}(t)\right) \tag{42}$$

The number of servers with packed requests, n' can be expressed as a function of n:

$$n' = \left\lceil \frac{nr(t)}{u_{max}(2.0)Ncap_{max}(f=2.0)} \right\rceil \leqslant n \tag{43}$$

By setting $P_{set}(t) = P_{DC}(t)$ and substituting (43) into (42), we can solve for n as follows:

$$n = \left\lceil \{P_{set}(t) - NP_i(f=2.1, u=u_{uni}\}/\left\{ \frac{r(t)P_i(f=2.0, u=u_{max}(2.0))}{u_{max}(2.0)Ncap_{max}(f=2.0)} + \left[1 - \frac{r(t)}{u_{max}(2.0)Ncap_{max}(f=2.0)}\right]P_0 - P_i\left(f=2.1, u=u_{uni}\right)\right\}\right\rceil \tag{44}$$

n' can then be derived from Eq. (43). As shown in the Lemma III of Appendix C, by increasing n, the total power consumption can be continuously reduced from $P_{uni}(t)$ to the power consumption lower bound in Eq. (C.12) with an error less than the power consumption of one server. The approximated minimum power consumption is reached when $n = N$, i.e. all workload are packed to servers running at 2.0 GHz.

CASE 3: When $P_{set}(t) < P_{uni}(t)$ and $r(t) > u_{max}(2.0) \times N \times cap_{max}(f=2.0)$, the data center's operating strategy works as follows. The combination of a low power set point and a large number of requests pushes the utilization rate of the servers to the upper limit. Note that even by running all servers at 2.0 GHz with the utilization rate at the upper limit which satisfies the SLA, the data center can only handle $N \times u_{max}(2.0) \times cap_{max}(f=2.0)$ requests per second. This is smaller than the number of $r(t)$ in CASE 3. Therefore, only n out of a total of N servers can operate at 2.0 GHz with a utilization rate of $u_{max}(2.0)$. The remaining $n(r(t)/N - u_{max}(2.0)cap_{max}(f=2.0))$ workload will be evenly distributed to the remaining $N - n$ servers operating at a frequency of 2.1 GHz.

The number of servers, n, operating at 2.0 GHz can be calculated online as follows.

The total energy consumption of the data center in CASE 3 include energy consumption from n servers at frequency 2.0 GHz and the utilization rate of $u_{max}(2.0)$, and $N - n$ servers operating at frequency 2.1 GHz with utilization rate of $u(t)$.

$$P_{DC}(t) = \sum_{i=1}^{n} P_i\left(f=2.0, u_{max}(2.0)\right) + \sum_{i=n+1}^{N} P_i\left(f=2.1, u(t)\right) \tag{45}$$

where

$$u(t) = \frac{r(t) - n \times u_{max}(2.0)cap_{max}(f=2.0)}{(N-n)cap_{max}(f=2.1)} \tag{46}$$

The utilization rate of the $N - n$ servers operating at 2.1 GHz satisfies the following relationship:

$$u(t) > u_{uni}(t) > \frac{u_{max}(2.0)cap_{max}(f=2.0)}{cap_{max}(f=2.1)} > 0.1 \tag{47}$$

Therefore, by setting $P_{set}(t) = P_{DC}(t)$ and combining Eqs. (45) and (46), the closed-form solution for n can be derived as:

$$n = \left\lfloor \left\{ N\left(2.1\alpha_3^4 + \alpha_4^4\right) + \frac{(2.1\alpha_1^4 + \alpha_2^4)r}{cap_{max}(f=2.1)} - P_{set}(t) \right\} \Big/ \left\{ 2.1\alpha_3^4 + \alpha_4^4 + \left(2.1\alpha_1^4 + \alpha_2^4\right)\frac{u_{max}(2.0)cap_{max}(f=2.0)}{cap_{max}(f=2.1)} - P_i\left(f=2.0, u=u_{max}(2.0)\right)\right\}\right\rfloor \tag{48}$$

As shown in the Lemma IV of Appendix C, by gradually increasing n, the total power consumption can be reduced from $P_{uni}(t)$ to the power consumption lower bound in Eq. (C.18) with an error less than the power consumption of two servers. The approximated minimum power consumption is reached when the $n = N - n^*_{2.1}$, where $n^*_{2.1}$ is defined in Eq. (C.1).

In summary, the rule-based data center power consumption control strategy is presented in Algorithm 1.

Algorithm 1. Rule-Based Data Center Power Consumption Control Strategy

1: Receives data center power set point $P_{set}(t) = P_{DC}^{base}(t) + B_{cap}(t)RegD(t)$.
2: Calculate total server power consumption with uniformly distributed requests at the highest CPU frequency as $P_{DC}(u_{uni}(t), f_{max}) = \sum_i^N P_i(f=2.1, u=u_{uni}(t))$.
3: **if** $P_{set}(t) \geqslant P_{DC}(u_{uni}(t), f_{max})$ **then**
4: Add dummy load evenly to all servers to increase utilization rate as in (38)–(41)
5: **else if** $r(t) < Nu_{max}(2.0)cap_{max}(f=2.0)$ **then**
6: Calculate n', the number of servers operating at frequency 2.0 GHz and utilization rate $u_{max}(2.0)$, and $n - n'$, the number of servers idling, as in (43) and (44). The remaining servers will be operating at frequency 2.1 GHz and utilization rate of $u = u_{uni}(t)$.
7: **else**
8: Calculate n, the number of servers operating at frequency 2.0 GHz and utilization rate $u_{max}(2.0)$ as in (48). The rest of the servers are operating at frequency 2.1 - GHz and utilization rate as in (46).
9: **end if**

6. Numerical study

6.1. Simulation setup

It is assumed that the data center in the numerical study has 100,000 servers. Given estimates of data center size from 2017 [37] ranging from 50,000–80,000, we believe that 100,000 servers in a data center is reasonable to simulate a large data center. The servers are assumed to have the same power curves as shown in Fig. 4 with power consumption ranging from 22 W to 85 W. The maximum capacity of each server is 1230 requests per second. The SLA specifies that 90% of the requests will be processed within 115 ms. The corresponding limit on the utilization rate are 0.8 at 2.1 GHz and 0.77 at 2.0 GHz. To simulate the data center's workload, we adopted Wikipedia's access trace from the online repository [38]. The historical prices for frequency regulation and energy from the PJM market are used for electricity market simulation. According to the data center efficiency reports [39,40], the average power usage effectiveness (PUE) of Google data centers is about 1.12, i.e., the non-IT load is about 12% of IT load. The PUEs of Google and Facebook data centers are as low as 1.07. Assuming that cooling is the major non-IT load, η_{cool} is set to be 0.12 in the following simulation.

6.2. Performance of data center requests forecast

The accuracy of the data center requests forecast is crucial to determining the optimal level of bidding quantity for both energy and frequency regulation service. In the numerical study, the Wikipedia access trace is adopted to simulate the data center's workload. Wikipedia is one of the most visited websites on the Internet, and it is hosted on more than 350 servers [41]. The Wikipedia's access trace contains about 10 percent of all users' requests to Wikipedia, which is collected over about 32 days with a granularity of a millisecond.

The requests for visiting English, Spanish and Polish web pages are used to evaluate the performance of the data center request forecast. The snapshots of the three groups of web page access data are shown in Fig. 6.

An ARIMA model is built to perform rolling hour-ahead data center workload trace prediction. The last week of the workload trace of

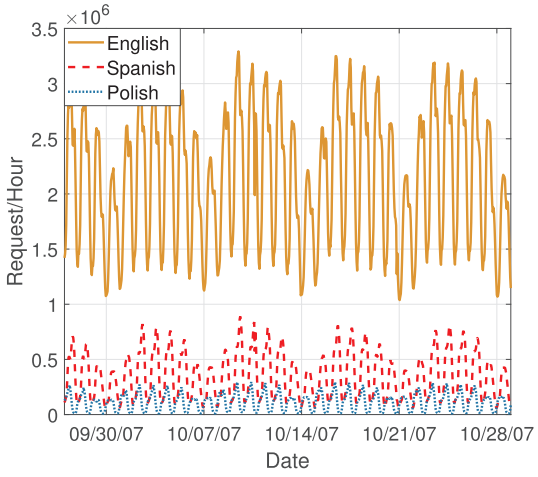

Fig. 6. Request traces of English, Spanish and Polish pages.

Wikipedia data is used for out-of-sample testing. The forecasted and actual workload trace are depicted in Fig. 7. As shown in Fig. 7, the hour-ahead workload trace prediction is fairly accurate. To quantify the performance of the forecast method, the mean absolute percentage error (MAPE) of the hour-ahead workload forecast and the standard deviation normalized over mean of the 5-min workload data of the last week are shown in Table 1. As shown in the table, the forecast accuracy for the English web page is the highest with a MAPE smaller than 5%. Because the normalized standard deviation of the number of visits for the Spanish and Polish pages are higher, the corresponding prediction errors are also larger.

In the following simulations, the requests arrival rate is scaled up so that the peak hourly utilization rate for the data center with 100,000 host servers is about 65% when the CPU operating frequency is 2.1 GHz.

6.3. Performance of the electricity price forecast

The net earnings for a data center to participate in the frequency regulation market depend on the difference between the frequency regulation service price and the energy price. When the frequency regulation service price is higher than the energy price, i.e., the price difference is positive, the data center receives extra benefits by providing frequency regulation services compared to simply minimizing its power consumption. Hence, the accuracy of the price difference prediction is crucial to the successful implementation of the frequency regulation provision strategy for data centers.

The historical prices for frequency regulation services and energy from the PJM market is leveraged to build and test the price difference forecasting algorithms. The explanatory variables in the price forecast model include the hourly electricity demand, hourly electricity generation by energy source, and hour of the day. The input features of the price forecast model are summarized in Table 2. The data sets from the last week of each month in 2017 are chosen as the test set. The remaining data in 2017 are used for training and validation with a split-ratio of 0.7/0.3. Feed-forward neural network with batch normalization and two hidden layers of 256 neurons and 128 neurons respectively is trained. Early stopping [42] technique is adopted based on the F1 score of the binary classification problem for predicting the sign of the price difference. F1 score is a widely used metric for binary classification. The positive class corresponds to the case where the frequency regulation price is larger than the effective energy price, while the negative class corresponds to the case where the regulation price is smaller than the effective energy price. The performance score of the signal following is assumed to be 100% here. Although the forecast for the magnitude of price difference is necessary for the risk estimation, the sign of the frequency regulation price minus the effective energy price is more important in terms of the total gain. As reported in Table 3, the F1 score is 0.70 in the test set. Note that, for the false negative cases, i.e., the expected price difference is wrongly predicted to be negative, the data center will operate in the power minimization mode, which does not incur extra cost.

6.4. Performance of frequency regulation service provision by data center

The performance of the frequency regulation service provision by data center will be evaluated from three perspectives: frequency regulation signal following performance, electricity cost, and request response time. The price prediction result of the 12 last weeks in each month of the year 2017 is used in the simulation. During the performance evaluation, we assume that the data center will provide frequency regulation service to the electricity market whenever the expected frequency regulation service price is higher than the energy price. In other words, we do not consider the risk constraint. In the real-time operation simulations, the data center is expected to follow the historical frequency regulation signals from the PJM market. The requests served by the data are derived from the scaled requests arrival rate of English, Spanish and Polish pages in the last week of Wikipedia trace as shown in Fig. 8, which is repeatedly used. The utilization rate of each server is determined by Algorithm 1 with the bi-linear power model. The actual power consumption of the data center is estimated with empirical measurement data with interpolation.

The frequency regulation signal following performance of the proposed data center power consumption control algorithm is quantified by three metrics: accuracy, delay, and precision. The frequency regulation signal and the actual power consumption trajectory of the data center for an hour is depicted in Fig. 9. It can be seen from the figure that the proposed data center power consumption control algorithm allows the data center to closely follow the frequency regulation signals. The accuracy and precision scores for the 12 weeks are calculated and shown in Table 4. The power set point can be determined by the rule-based control algorithm within about 0.2 s with a MATLAB program on a standard Dell desktop computer. The delay scores are all 100% for all the test cases. It can be seen that the accuracy and precision

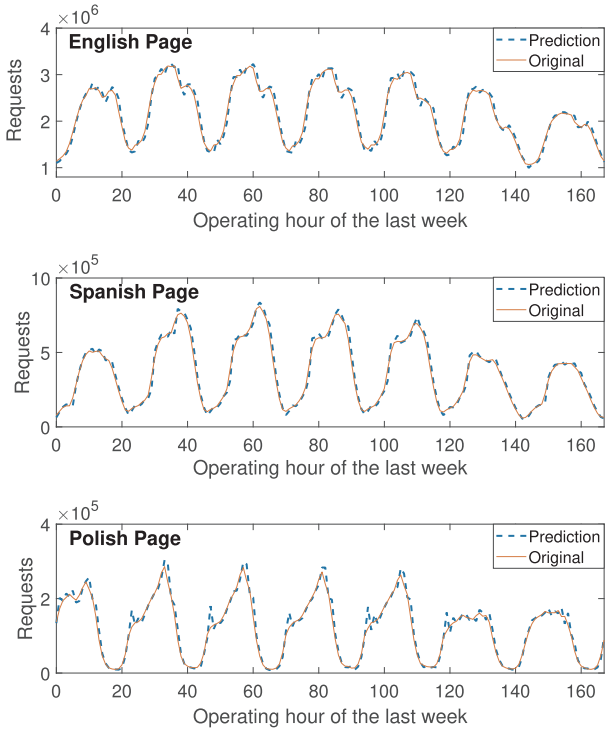

Fig. 7. Prediction of English, Spanish and Polish page visits.

Table 1
Forecast performance comparison among English, Spanish, and Polish Web Pages.

	English Page	Spanish Page	Polish Page
MAPE	4.09%	7.37%	11.70%
Normalized Standard Deviation	3.58%	10.28%	21.79%

Table 2
Extracted features.

Features	
Last-4-h Prices	Energy Prices
	Regulation Capacity Clearing Prices
	Performance Clearing Prices
	Mileage Ratio
Last-4-h Generation	Solar Generation
	Wind Generation
	Storage Generation
	Hydro Generation
	Other Renewable
	Nuclear Generation
	Coal Generation
	Oil Generation
	Gas Generation
	Multiple Fuels
	Other Generation
Last-4-h Load	Total Demand
Time	Operating Hour of the Day

Table 3
Performance of the price difference forecast.

	Training	Validation	Testing
F1 score	0.73	0.68	0.70

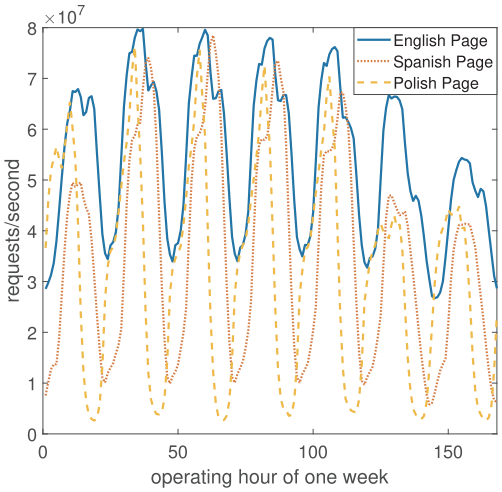

Fig. 8. Hourly-averaged request arrival rate after scaling.

scores of the data center are very high for the three types of page visit traces. The small frequency regulation signal tracking errors mainly come from two sources: the requests prediction error and the approximation error of the piece-wise bi-linear server power model. For comparison, the signal following performance of the benchmarks without dummy load and using a linear power model [19] are also provided. When the dummy load control knob is removed, the data center is unable to closely follow the frequency regulation signal when

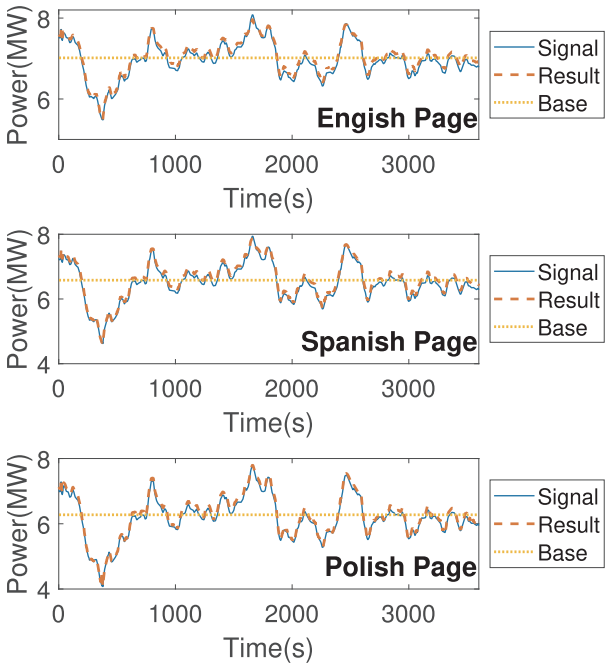

Fig. 9. Frequency regulation signal following of one hour for the three different pages.

the power set point is higher than the maximum power consumption under uniform request routing. Thus, the introduction of dummy load increases the range of power consumption of the data center. This allows the data center to meet the power set points that frequency scaling alone cannot meet. In terms of using a linear power model, the accuracy score remains reasonable as the correlation is still high. However, the large power estimation error of the linear model results in a significant drop in precision score.

We now present the reduction in electricity cost by participating in the frequency regulation market for the data center. If the data center does not provide frequency regulation service to the power system, then it will operate to minimize its power consumption. The electricity costs of the data center with the proposed algorithm and benchmarks including removing the control knob of the dummy load and the power minimization strategy are reported in Table 5. For a data center with 100,000 servers, the proposed data center control algorithm results in a $21,590 (8.1%) electricity costs reduction for the 12 weeks on average compared to the power minimization strategy. With the introduction of dummy load, a higher upper bound of feasible power consumption can be achieved, which improves the cost saving of frequency regulation by about 300% compared to the case without dummy load as shown in Table 5. Moreover, the revenue from providing frequency regulation services with the linear power model cannot cover the increased electricity cost due to poor signal following performance. Hence, the introduction of dummy load and the adoption of a bi-linear model are crucial to the profitability of the data center frequency regulation service provision.

The requests response time of the data center when providing frequency regulation services is calculated based on the proposed request routing algorithm. The distribution of request response time during the hours when the data center provides frequency regulation is shown in Fig. 10. Compared to the uniform request routing strategy, when the data center follows the frequency regulation signals, only a small portion of requests' response time moved closer to the SLA's response time limit. If the data center does not provide frequency regulation service and instead minimizes power consumption with a packing strategy, then the response time of almost all the requests will be very close to the SLA's response time limit. Hence, compared to the minimum power

Table 4
Frequency regulation signal following performance scores.

Dummy load	Power model	Performance score	English page	Spanish page	Polish page
Yes	Bilinear	Accuracy	99.76%	99.69%	99.62%
		Precision	95.35%	95.87%	95.74%
	Linear	Accuracy	96.37%	96.89%	96.28%
		Precision	52.40%	56.99%	55.38%
No	Bilinear	Accuracy	99.58%	98.87%	95.44%
		Precision	92.59%	92.54%	92.36%
	Linear	Accuracy	43.97%	52.45%	42.21%
		Precision	8.08%	20.04%	20.34%

consumption control strategy, the proposed data center control with frequency regulation provision reduces not only electricity costs, but also the response time of requests.

6.5. Impacts of risk limit on frequency regulation bidding capacity and net earnings

The impacts of risk limit on frequency regulation bidding capacity and the data center's net earnings are evaluated in this subsection. As shown in the earlier subsection, the performance scores of the data center in following frequency regulation signals are almost perfect, hence the performance scores are assumed to be 100% in the evaluation here. For illustrative purpose, only the scaled English page traces are used in the simulation of this subsection. By setting the weekly bidding risk limit δ_{risk} at \$336, the frequency regulation bidding quantities and the differences between frequency regulation prices and energy prices are shown in Fig. 11 for one sample week. In the figure, the green dashed line represents the maximum feasible frequency regulation bidding capacity calculated based on the predicted requests arrival rate of the next hour. The red squares denote the predicted hourly price differences, while the blue dots are the frequency regulation bidding capacity obtained from the risk-limited data center bidding strategy. It can be seen from the figure that the bidding capacities are zero when the expected price differences are negative. In addition, the actual bidding capacity is scaled down from the maximum feasible bidding capacity when the confidence level in the positive price difference forecast is low. Next, we gradually increased the weekly bidding risk limit from \$0 to \$3,000 and recorded the extra earnings of the data center by providing frequency regulation services compared to the benchmark case where the data center minimizes power consumption. The trade-off between the weekly risk limit and the extra net earnings are depicted in Fig. 12. As shown in the figure, the extra net earnings made by the data center increases with the bidding risk limit. However, when the risk limit is very high, the saturation effect kicks in, which leads to a slow increase in extra net earnings with the risk limit.

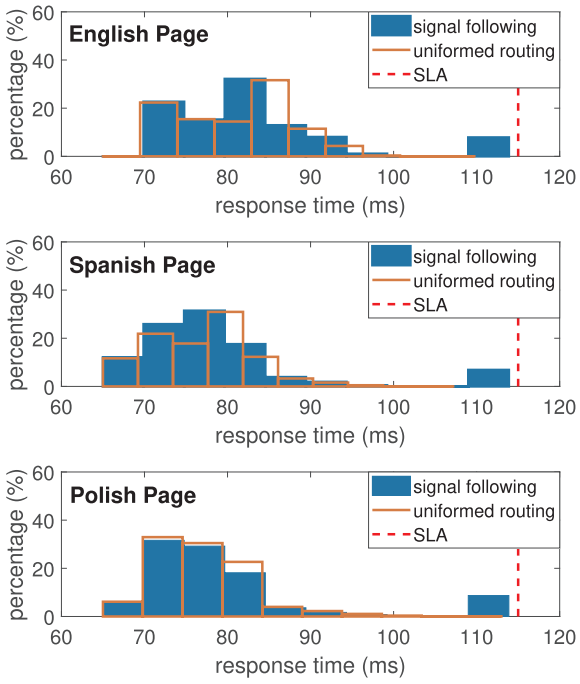

Fig. 10. Distribution of requests response time.

7. Conclusion

The operational flexibility of the data centers can be leveraged to provide valuable frequency regulation services in the smart grid. A comprehensive frequency regulation service provision framework is proposed in the paper. A risk constrained hour-ahead bidding strategy and a real-time frequency regulation signal following algorithm are developed. The introduction of dummy load and the realistic server power consumption model allow data centers to follow real-world frequency regulation signals with over 95% accuracy. Numerical study with Wikipedia's access trace shows that with reliable energy and frequency regulation service price forecast, data centers can reduce their electricity bill by more than 8% without violating service level agreements.

Acknowledgement

This work was supported by the Department of Energy under award (#DE-OE0000840) and the California Energy Commission under award (EPC-16-030).

Table 5
Electricity costs of the data center.

	Dummy load	Power model	English Page	Spanish Page	Polish Page
Cost With Frequency Regulation (\$)	Yes	Bilinear	298.38 K	238.29 K	217.91 K
		Linear	326.83 K	271.15 K	251.08 K
	No	Bilinear	309.92 K	254.31 K	233.49 K
		Linear	337.16 K	278.01 K	252.41 K
Costs with Minimum Power(\$)			316.65 K	262.62 K	240.08 K

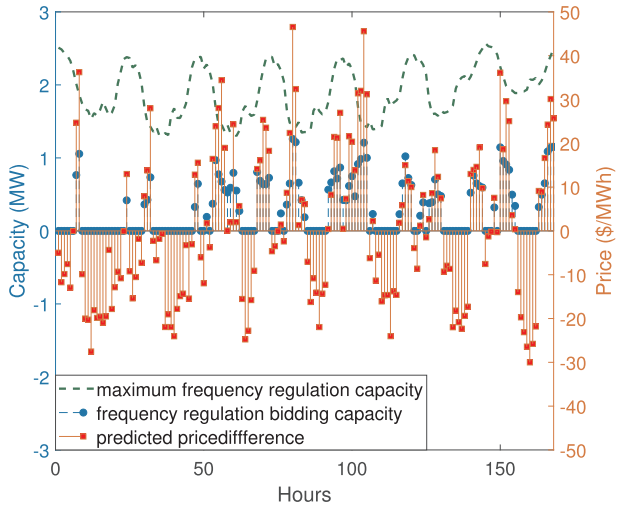

Fig. 11. Frequency regulation service bidding capacity versus predicted price difference with weekly risk limit of $336.

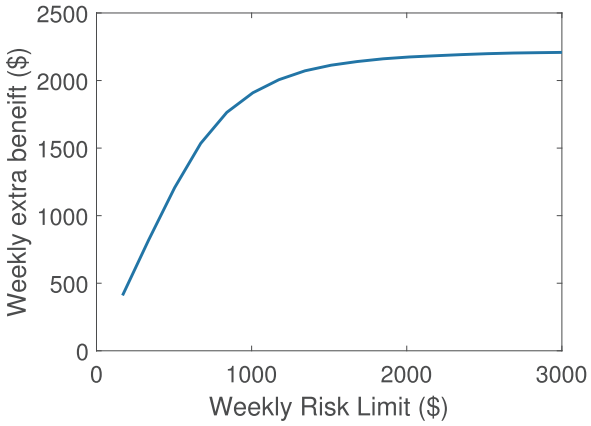

Fig. 12. Weekly extra net earnings versus the risk limit.

Appendix A. Derivative of risk function

Lemma 1. *If the frequency regulation service price is positive, then the derivative of risk function with respect to power consumption base is greater than or equal to zero.*

Proof. For any $\Delta P > 0$,

$$f_{risk}(P_{DC}^{base}(t) + \Delta P)$$

$$= \iint_{V'} Pr(C_{reg}(t), C_{efe}(t))\{C_{efe}(t)[P_{DC}^{base}(t) + \Delta P - P_{DC}^{min}(t)] - C_{reg}(t)score(t)B_{cap}(t)\}dC_{reg}(t)dC_{efe}(t) \tag{A.1}$$

where the integral region is

$$V' = \{C_{reg}(t), C_{efe}(t) | C_{efe}(t)[P_{DC}^{base}(t) + \Delta P - P_{DC}^{min}(t)] > C_{reg}(t)score(t)B_{cap}(t)\} \tag{A.2}$$

when $P_{DC}^{base}(t) = P_{DC}^{min}(t)$, we have $B_{cap}(t) = 0$. This is because $B_{cap}(t) \geqslant 0$ and $P_{DC}^{base}(t) - B_{cap}(t) \geqslant P_{DC}^{min}(t)$.

In this case,

$$f_{risk}(P_{DC}^{base}(t) + \Delta P) \geqslant f_{risk}(P_{DC}^{base}(t)) = 0 \tag{A.3}$$

When $P_{DC}^{base}(t) > P_{DC}^{min}(t)$,

$$V = \left\{ C_{reg}(t), C_{efe}(t) \middle| C_{efe}(t) > \frac{C_{reg}(t)score(t)B_{cap}(t)}{P_{DC}^{base}(t) - P_{DC}^{min}(t)} \right\} \tag{A.4}$$

$$V' = \left\{ C_{reg}(t), C_{efe}(t) \middle| C_{efe}(t) > \frac{C_{reg}(t)score(t)B_{cap}(t)}{P_{DC}^{base}(t) + \Delta P - P_{DC}^{min}(t)} \right\} \tag{A.5}$$

We can see that, with $C_{reg}(t) > 0$, $V \subseteq V'$. Therefore, we have

$$f_{risk}\left(P_{DC}^{base}(t)+\Delta P\right)-f_{risk}\left(P_{DC}^{base}(t)\right)=\iint_{V'-V}Pr\left(C_{reg}(t),\,C_{efe}(t)\right)\{C_{efe}(t)[P_{DC}^{base}(t)+\Delta P-P_{DC}^{min}(t)]-C_{reg}(t)score(t)B_{cap}(t)\}dC_{reg}(t)$$
$$dC_{efe}(t)+\iint_{V}Pr\left(C_{reg}(t),\,C_{efe}(t)\right)C_{efe}(t)\Delta P dC_{reg}(t)dC_{efe}(t)\geqslant 0 \tag{A.6}$$

Hence,

$$\frac{\partial f_{risk}}{\partial P_{DC}^{base}(t)}=\lim_{\Delta P\to 0}\frac{f_{risk}(P_{DC}^{base}(t)+\Delta P)-f_{risk}(P_{DC}^{base}(t))}{\Delta P}\geqslant 0 \tag{A.7}$$

Appendix B. Proof of the minimum response time

Lemma 2. *The minimum total request response time is achieved when the workload is uniformly distributed to all servers running at 2.1 GHz.*

Proof. The response time decreases with the increase of frequency for a given utilization rate. The utilization rate also decreases with the increase of frequency for a given amount of requests. Hence, the response time decreases with the increase of frequency at each server.

$$\sum_{i=1}^{N}r_i(t)rt_i(t)=\sum_{i=1}^{N}r_i(t)f_{rt}\left(f_i(t),\frac{r_i(t)}{cap_{max}(f_i(t))}\right)\geqslant\sum_{i=1}^{N}r_i(t)f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right) \tag{B.1}$$

Ignoring the frequency regulation signal following constraint, the total request response time minimization problem can be reformulated as

$$\min\sum_{i=1}^{N}r_i(t)f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right) \tag{B.2}$$

s.t.

$$\sum_{i=1}^{N}r_i(t)=r(t) \tag{B.3}$$

Define $f_{rt,2.1}=f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)$. Then it can be seen from Fig. 5 that

$$\frac{d\,f_{rt,2.1}}{d\,r_i(t)}=\frac{1}{cap_{max}(f_{max})}\frac{d\,f_{rt,2.1}}{d\,u_i(t)}\geqslant 0 \tag{B.4}$$

$$\frac{d^2f_{rt,2.1}}{(d\,r_i(t))^2}=\frac{1}{(cap_{max}(f_{max}))^2}\frac{d^2f_{rt,2.1}}{(d\,u_i(t))^2}\geqslant 0 \tag{B.5}$$

Thus,

$$\frac{d^2r_i(t)f_{rt,2.1}}{(d\,r_i(t))^2}=2\frac{d\,f_{rt,2.1}}{d\,r_i(t)}+r_i(t)\frac{d^2f_{rt,2.1}}{(d\,r_i(t))^2}\geqslant 0 \tag{B.6}$$

Therefore, the objective function (B.2), which is the sum of $r_i(t)f_{rt,2.1}$, is convex.

The Lagrange function of problem (B.2) and (B.3) is:

$$\mathscr{L}=\sum_{i=1}^{N}r_i(t)f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)+\lambda\left(r(t)-\sum_{i=1}^{N}r_i(t)\right) \tag{B.7}$$

where λ is the Lagrange multiplier.

By taking partial derivative with respect to $r_i(t)$,

$$\frac{\partial\mathscr{L}}{\partial r_i(t)}=f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)+r_i(t)\frac{\partial f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)}{\partial r_i(t)}-\lambda \tag{B.8}$$

At optimal solutions, (B.8) equals zero. Hence, we have

$$f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)+r_i(t)$$
$$\frac{\partial f_{rt}\left(f=2.1,\frac{r_i(t)}{cap_{max}(f_{max})}\right)}{\partial r_i(t)}=\lambda,\quad\forall\, i \tag{B.9}$$

The above optimality condition is achieved by uniformed routing, i.e. $r_i(t)=r_j(t),\ \forall\, i,j$.

Appendix C. Packing strategy monotonically reduces power consumption to the minimum power consumption

Lemma 3. *When* $r(t) \leqslant u_{max}(2.0) \times N \times cap_{max}(2.0)$, *starting from the baseline operating strategy where the requests are uniformly routed to all servers running at* $f = 2.1$ *GHz, the total power consumption can be reduced by selecting n servers whose workload are packed to* n' *servers running at 2.0 GHz with the maximum utilization rate* $u_{max}(2.0)$ *as in Eq. (42). By gradually increasing n, the power consumption can be monotonically reduced to the minimum power consumption with error less than the power consumption of one server, which is achieved when* $n = N$.

Lemma 4. *When* $r(t) > u_{max}(2.0) \times N \times cap_{max}(2.0)$, *starting from the baseline operating strategy where the requests are uniformly routed to all servers running at* $f = 2.1$ *GHz, the total power consumption can be reduced by curtailing workload on n servers to operate at 2.0 GHz with the maximum utilization rate* $u_{max}(2.0)$ *and uniformly distributing the remaining workload on the* $N - n$ *servers as in Eq. (45). By gradually increasing n, the power consumption can be monotonically reduced to the minimum power consumption with error less than the power consumption of two servers, which is achieved when* $n = N - n_{2.1}^*$, *where*

$$n_{2.1}^* = \frac{r(t) - u_{max}(2.0) N cap_{max}(2.0)}{u_{max}(2.1) cap_{max}(2.1) - u_{max}(2.0) cap_{max}(2.0)} \tag{C.1}$$

Proof. The total power consumption of a data center is:

$$P_{DC}(t) = \sum_{i=1}^{N} P_i(t) = \sum_{i=1}^{N} \left(P_i(t) - P_0 \right) + \sum_{i=1}^{N} (P_0) \tag{C.2}$$

The dynamic power consumption per request per second for server i is defined as

$$dpr_i(t) = \frac{P_i(t) - P_0}{r_i(t)} \tag{C.3}$$

Substitute the energy consumption model (4) into (C.3):

$$\begin{aligned} dpr_i(t) &= \frac{\alpha_1^j f_i(t) u_i(t) + \alpha_2^j u_i(t) + \alpha_3^j f_i(t) + \alpha_4^j - P_0}{r_i(t)} \\ &= \frac{\alpha_1^j f_i(t) u_i(t) + \alpha_2^j u_i(t) + \alpha_3^j f_i(t) + \alpha_4^j - P_0}{cap_{max}(f_i(t)) u_i(t)} \\ &= \frac{\alpha_1^j f_i(t) u_i(t) + \alpha_2^j u_i(t) + \alpha_3^j f_i(t) + \alpha_4^j - P_0}{\frac{cap_{max}}{f_{max}} f_i(t) u_i(t)} \end{aligned} \tag{C.4}$$

Eq. (C.3) can be transformed into the following form:

$$dpr_i(t) = \frac{\alpha_1^j}{cap_f} + \frac{\alpha_2^j}{cap_f f_i(t)} + \frac{\alpha_3^j + \frac{\alpha_4^j - P_0}{f_i(t)}}{cap_f u_i(t)} \tag{C.5}$$

where $cap_f = cap_{max}/f_{max}$. By plugging $u_i(t) = 0$ into Eq. (4), we get $\alpha_3^j f_i(t) + \alpha_4^j - P_0 \geqslant 0$ or equivalently:

$$\alpha_3^j + \frac{\alpha_4^j - P_0}{f_i(t)} \geqslant 0 \tag{C.6}$$

Therefore, for a given CPU frequency, the first two terms on the right-hand side of (C.5) are constant, and the last term decreases with the increase of $u_i(t)$. In other words, the dynamic power consumption per request for server i, $dpr_i(t)$, decreases with higher utilization rate.

It can be seen in Fig. C.13 that for a given utilization rate, the dynamic power consumption per request, dpr_i, roughly stays at the same value when $1.2 \leqslant f \leqslant 2.0$. However, as shown in Fig. 5, for a given SLA limit rt_{SLA}, a higher utilization rate can be reached when the CPU frequency increases from 1.2 GHz to 2.0 GHz. In other words, $u_{max}(f = 2.0) > u_{max}(f)$, $\forall f \in [1.2, 2.0)$. As the $dpr_i(t)$ decreases with higher $u_i(t)$, $dpr_i(f = 2.0, u = u_{max}(f = 2.0)) < dpr_i(f, u = u_{max}(f))$, $\forall f \in [1.2, 2.0)$.

It can also be seen in Fig. C.13 that dpr_i jumps when f increases from 2.0 GHz to 2.1 GHz. By performing a calculation with server power consumption curve and dynamic power curve parameters, we can verify that $dpr_i(f = 2.1, u = u_{max}(f = 2.1)) > dpr_i(f = 2.0, u = u_{max}(f = 2.0))$. Therefore, the minimum dynamic power consumption per request of each server, $dpr_i^{min}(t)$, is achieved when the requests are packed to fully utilize

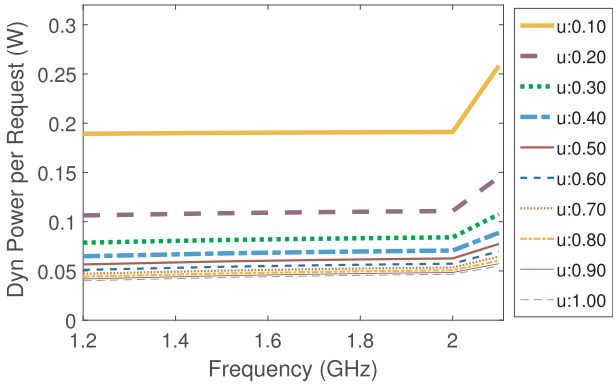

Fig. C.13. Dynamic power consumption per request.

the servers operating with CPU frequency of 2.0 GHz.

For Lemma 3: If the total requests arrival rate falls into the following range, $r(t) \leqslant u_{max}(2.0) \times N \times cap_{max}(2.0)$, we can pack the workload of n servers to n' servers with $f_i(t) = 2.0$ and $u_i(t) = u_{max}(2.0)$ according to Eq. (43).

Let's select two different numbers of servers n_1 and n_2 for packing, where $n_1 < n_2$. We will be packing the workload of n_1 (n_2) servers to n_1' (n_2') servers running at 2.0 GHz with a utilization rate of $u_{max}(2.0)$. n_1' and n_2' can be calculated as follows:

$$n_1' = \left\lceil \frac{n_1 r(t)}{u_{max}(2.0) N cap_{max}(f = 2.0)} \right\rceil \tag{C.7}$$

$$n_2' = \left\lceil \frac{n_2 r(t)}{u_{max}(2.0) N cap_{max}(f = 2.0)} \right\rceil \tag{C.8}$$

Thus, $n_1' \leqslant n_2'$. Let's define the workload handled by he n_1' and n_2' servers as

$$r_1(t) = n_1' u_{max}(2.0) cap_{max}(2.0) \tag{C.9}$$

$$r_2(t) = n_2' u_{max}(2.0) cap_{max}(2.0) \tag{C.10}$$

Since $n_1' \leqslant n_2'$, we have $r_1(t) \leqslant r_2(t)$.

Because $dpr_i(f = 2.1, u = u_{uni}(t)) > dpr_i^{min}(t)$, we can show that

$$\begin{aligned} P_{uni}(t) &= \sum_{i=1}^{n} \left(P_i(t) - P_0 \right) + \sum_{i=1}^{n} P_0 \\ &= dpr_i\left(f = 2.1, u = u_{uni}(t)\right) r(t) + \sum_{i=1}^{N} P_0 \\ &\geqslant dpr_i^{min}(t) r_1(t) + dpr_i(f = 2.1, u = u_{uni}(t)) \\ &\qquad \left[r(t) - r_1(t) \right] + \sum_{i=1}^{N} P_0 \\ &\geqslant dpr_i^{min}(t) r_2(t) + dpr_i(f = 2.1, u = u_{uni}(t)) \\ &\qquad \left[r(t) - r_2(t) \right] + \sum_{i=1}^{N} P_0 \end{aligned} \tag{C.11}$$

Therefore, by gradually increasing n, the power consumption reduces monotonically.

For any feasible request routing strategy, we have

$$\begin{aligned} P_{DC}(t) &= \sum_{i=1}^{N} P_i(t) = \sum_{i=1}^{N} \left(P_i(t) - P_0 \right) + \sum_{i=1}^{N} P_0 \\ &= \sum_{i=1}^{N} dpr_i(t) r_i(t) + \sum_{i=1}^{N} P_0 \\ &\geqslant dpr_i^{min}(t) \sum_{i=1}^{n_{2.0}} r_i(t) + \sum_{i=1}^{N} P_0 \\ &= dpr_i^{min}(t) r(t) + \sum_{i=1}^{N} P_0 = \underline{P}_1 \end{aligned} \tag{C.12}$$

where $n_{2.0}$ is the number of servers running at 2.0 GHz with a utilization rate of $u_{max}(2.0)$ after packing all the workload.

$$n_{2.0} = \left\lceil \frac{r(t)}{u_{max}(2.0) N cap_{max}(2.0)} \right\rceil \tag{C.13}$$

For Lemma 4: If the total requests arrival rate falls into the following range, $r(t) > u_{max}(2.0) \times N \times cap_{max}(2.0)$, to satisfy the SLA constraint, there must be some servers running at $f = 2.1$ GHz to handle the additional workload.

In this case, the data center power consumption can be reduced by decreasing the workload on n servers by reducing the CPU frequency from 2.1 GHz to 2.0 GHz with the maximum utilization rate $u_{max}(2.0)$. The remaining workload will be uniformly distributed to the $N - n$ servers according to Eq. (45).

Let's select two different number of servers n_1 and n_2 for workload and CPU frequency reduction, where $n_1 < n_2$.

Define the workload handled by the n_1 and n_2 servers as r_1 and r_2. The workload can be calculated as

$$\begin{aligned} r_1(t) &= n_1 u_{max}(2.0) cap_{max}(2.0) \\ r_2(t) &= n_2 u_{max}(2.0) cap_{max}(2.0) \end{aligned} \tag{C.14}$$

Because $n_1 < n_2$, we have $r_1(t) < r_2(t)$.

Now, let's define the utilization rate of the remaining $N - n_1$ and $N - n_2$ servers as u_1 and u_2. The utilization rates can be calculated as

$$u_1(t) = \frac{n_1(u_{uni}(t) cap_{max}(2.1) - u_{max}(2.0) cap_{max}(2.0))}{N - n_1} + u_{uni}(t)$$

$$u_2(t) = \frac{n_2(u_{uni}(t)cap_{max}(2.1) - u_{max}(2.0)cap_{max}(2.0))}{N - n_2} + u_{uni}(t) \tag{C.15}$$

Therefore, $u_1(t) < u_2(t)$ and $dpr_i(f = 2.1, u = u_1(t)) > dpr_i(f = 2.1, u = u_2(t))$.

Now it can be shown that

$$\begin{aligned} P_{uni}(t) &= dpr_i\left(f = 2.1, u_{uni}(t)\right)r(t) + \sum_{i=1}^{N} P_0 \\ &\geqslant dpr_i^{min}(t)r_1(t) + dpr_i(f = 2.1, u_1(t)) \\ &\left[r(t) - r_1(t)\right] + \sum_{i=1}^{N} P_0 \\ &\geqslant dpr_i^{min}(t)r_2(t) + dpr_i(f + 2.1, u_2(t))(t)) \\ &\left[r(t) - r_2(t)\right] + \sum_{i=1}^{N} P_0 \end{aligned} \tag{C.16}$$

Therefore, by gradually increasing n, the data center power consumption will decrease monotonically.

When $n = N - n_{2.1}^*$, the minimum power consumption is achieved, where

$$n_{2.1}^* = \left\lceil \frac{r(t) - u_{max}(2.0)Ncap_{max}(2.0)}{u_{max}(2.1)cap_{max}(2.1) - u_{max}(2.0)cap_{max}(2.0)} \right\rceil \tag{C.17}$$

Note that $n_{2.1}^*$ is the minimum number of servers that have to operate at 2.1 GHz.

Denote the minimum dynamic power consumption per request $dpr_i(t)$ at $f = 2.1$ as $dpr_i^{min}(t, f = 2.1)$.

Then, we have

$$P_{DC}(t) \geqslant dpr_i^{min}(t) \sum_{i=n_{2.1}^*+1}^{N} r_i(t) + \sum_{i=1}^{N} P_0 + dpr_i^{min}\left(t, f = 2.1\right) \sum_{i=1}^{n_{2.1}^*} r_i(t) = \underline{P}_2 \tag{C.18}$$

Next we will prove that Eq. (C.18) holds and $\underline{P}_2$ is the lower bound of power consumption when $r(t) \geqslant u_{max}(2.0) \times N \times cap_{max}(2.0)$.

Let's denote the workload handled by the servers running at $f = 2.0$GHz as $r'(t)$. $r'(t)$ can then be calculated as

$$r'(t) = \sum_{i=n_{2.1}^*+1}^{N} r_i(t) = \left(N - n_{2.1}^*\right)u_{max}(2.0)cap_{max}(2.0) \tag{C.19}$$

Let's denote the remaining workload to be handled by servers running at $f = 2.1$ as $r^*(t)$. Thus we have

$$r^*(t) = r(t) - r'(t) \tag{C.20}$$

Denote the average dynamic power consumption per request of $r^*(t)$ as $dpr(r^*(t))$ and the average dynamic power consumption per request of $r'(t)$ as $dpr(r'(t))$. As $dpr(r^*(t)) \geqslant dpr_i^{min}(t, f = 2.1)$ and $dpr(r'(t)) \geqslant dpr_i^{min}(t)$, then for any feasible operating point, we have

$$P_{DC}(t) = r^*(t)dpr\left(r^*(t)\right) + r'(t)dpr\left(r'(t)\right) + \sum_{i=1}^{N} P_0 \geqslant dpr_i^{min}\left(t, f = 2.1\right)r^*(t) + dpr_i^{min}(t)r'(t) + \sum_{i=1}^{N} P_0 \tag{C.21}$$

Appendix D. Proof of the minimum amount of requests with increased response time

Lemma 5. *The amount of requests with increased response time compared to the uniform routing is minimized with the packing strategy in the CASE 2 ($P_{set}(t) < P_{uni}(t)$ and $r(t) \leqslant u_{max}(2.0) \times N \times cap_{max}(f = 2.0)$) of Section 5.2.*

Proof. For the real-time control strategy of CASE 2 in section 5.2, the response time of the requests on the $N - n$ servers running at $f = 2.1$ GHz is the same as that of the uniformed routing. In order to reduce data center power consumption from the power consumption under uniform routing policy $P_{uni}(t)$ by $\Delta P(t)$, we have to increase the response time for $\Delta r(t)$ requests per second by reducing their dynamic power consumption from $dpr_{uni}(t)$ to $dpr_i(t)'$.

Note that $\Delta r(t)$, $\Delta P(t)$, $dpr_i(t)$, and $dpr_i(t)'$ satisfy the following relationship:

$$\Delta r(t) = \frac{\Delta P(t)}{dpr_{uni}(t) - dpr_i(t)'} \tag{D.1}$$

In order to minimize the number of requests with increased response time per second, $\Delta r(t)$, we have to set $dpr_i(t)'$ at its minimum which is $dpr_i^{min}(t)$.

Now, as shown in the Appendix C, the minimum dynamic power consumption per request per second can be achieved when the servers are operating at $f = 2.0$ GHz and $u_{max}(2.0)$. This is the same packing strategy for Case 2 in Section 5.2. Therefore, the packing strategy proposed for Case 2 minimizes the number of requests with increased response time.

Appendix E. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apenergy.2019.05.107.

References

[1] Shehabi A, Smith SJ, Sartor DA, Brown RE, Herrlin M, Koomey JG, et al. United states data center energy usage report. Tech. rep. Lawrence Berkeley National Laboratory; June 2016.
[2] Koomey J, Turner P, Stanley J, Taylor B. A simple model for determining true total cost of ownership for data centers. White Paper, Uptime Institute; January 2007.
[3] Xu R, Melhem R, Mossé D. A unified practical approach to stochastic dvs scheduling. Proceedings of the 7th ACM international conference on embedded software. 2007. p. 37–46.
[4] Herbert S, Marculescu D. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. Proceedings of the 2007 international symposium on low power electronics and design. 2007. p. 38–43.
[5] Li K. Power and performance management for parallel computations in clouds and data centers. J Comput Syst Sci 2016;82(2):174–90.
[6] Mahmud AH, He Y, Ren S. Bats: budget-constrained autoscaling for cloud performance optimization. In: The 2014 ACM international conference on measurement and modeling of computer systems; 2014. p. 563–64.
[7] Verma A, Ahuja P, Neogi A. Pmapper: power and migration cost aware application placement in virtualized systems. Proceedings of the 9th ACM/IFIP/USENIX international conference on middleware. 2008. p. 243–64.
[8] Lin M, Liu Z, Wierman A, Andrew LLH. Online algorithms for geographical load balancing. In: 2012 international green computing conference (IGCC); 2012. p. 1–10.
[9] Cho J, Kim Y. Improving energy efficiency of dedicated cooling system and its contribution towards meeting an energy-optimized data center. Appl Energy 2016;165:967–82.
[10] Ham S-W, Kim M-H, Choi B-N, Jeong J-W. Energy saving potential of various air-side economizers in a modular data center. Appl Energy 2015;138:258–75.
[11] Khalaj AH, Scherer T, Siriwardana J, Halgamuge SK. Multi-objective efficiency enhancement using workload spreading in an operational data center. Appl Energy 2015;138:432–44.
[12] Khalaj AH, Halgamuge SK. A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system. Appl Energy 2017;205:1165–88.
[13] Wiser RH, Bolinger M, 2015 wind technologies market report. Tech rep. Department of Energy; August 2016.
[14] Margolis R, Feldman D, Boff D. Q4 2016/Q1 2017 solar industry update. Tech rep. Department of Energy; April 2017.
[15] California ISO. Q3 2017 report on market issues and performance; December 2017.
[16] Ghamkhari M, Mohsenian-Rad H. Data centers to offer ancillary services. In: IEEE international conference on smart grid communications; 2012. p. 436–41.
[17] Shi Y, Xu B, Zhang B, Wang D. Leveraging energy storage to optimize data center electricity cost in emerging power markets. ACM e-Energy 2016:18:1–18:13.
[18] Li S, Brocanelli M, Zhang W, Wang X. Data center power control for frequency regulation. In: IEEE power energy society general meeting; 2013. p. 1–5.
[19] Chen H, Coskun AK, Caramanis MC. Real-time power control of data centers for providing regulation service. In: IEEE conference on decision and control; 2013. p. 4314–21.
[20] Chen H, Caramanis MC, Coskun AK. Reducing the data center electricity costs through participation in smart grid programs. In: International green computing conference; 2014. p. 1–10.
[21] Chen H, Caramanis MC, Coskun AK. The data center as a grid load stabilizer. In: Asia and South Pacific design automation conference; 2014. p. 105–12.
[22] Chen H, Zhang B, Caramanis MC, Coskun AK. Data center optimal regulation service reserve provision with explicit modeling of quality of service dynamics. In: IEEE conference on decision and control; 2015. p. 7207–13.
[23] Aksanli B, Rosing T. Providing regulation services and managing data center peak power budgets. In: Design, automation test in europe conference exhibition; 2014. p. 1–4.
[24] Alaper I, Honkapuro S, Paananen J. Data centers as a source of dynamic flexibility in smart girds. Appl Energy 2018;229:69–79.
[25] Li S, Brocanelli M, Zhang W, Wang X. Integrated power management of data centers and electric vehicles for energy and regulation market participation. IEEE Trans Smart Grid 2014;5(5):2283–94.
[26] Brocanelli M, Li S, Wang X, Zhang W. Joint management of data centers and electric vehicles for maximized regulation profits. In: 2013 international green computing conference proceedings; 2013. p. 1–10.
[27] Yu N, Liu C-C, Price J. Evaluation of market rules using a multi-agent system method. IEEE Trans Power Syst 2010;25(1):470–9.
[28] Bergen A. Power systems analysis, Prentice-Hall series in electrical and computer engineering. Prentice-Hall; 1986.
[29] PJM. Energy & ancillary services market operations. < http://www.pjm.com//media/documents/manuals/m11.ashx > .
[30] Fan X, Weber W-D, Barroso LA. Power provisioning for a warehouse-sized computer. Proceedings of the 34th annual international symposium on computer architecture, ISCA '07. 2007. p. 13–23.
[31] Wong D, Annavaram M. Knightshift: scaling the energy proportionality wall through server-level heterogeneity. In: 2012 45th annual IEEE/ACM international symposium on microarchitecture (MICRO); 2012. p. 119–30.
[32] Wong D, Annavaram M. Implications of high energy proportional servers on cluster-wide energy proportionality. In: 2014 IEEE 20th international symposium on high performance computer architecture (HPCA); 2014.
[33] Ferdman M, Adileh A, Kocberber O, Volos S, Alisafaee M, Jevdjic D, et al. Clearing the clouds: a study of emerging scale-out workloads on modern hardware. In: Proceedings of the 17th international conference on architectural support for programming languages and operating systems; 2012. p. 37–48.
[34] Weaver VM, Johnson M, Kasichayanula K, Ralph J, Luszczek P, Terpstra D, et al. Measuring energy and power with PAPI. Proceedings of the 41st international conference on parallel processing workshops. 2012. p. 262–8.
[35] Wong D. Peak efficiency aware scheduling for highly energy proportional servers. In: Proceedings of the 43th annual international symposium on computer architecture; 2016.
[36] Brockwell PJ. Introduction to time series and forecasting. New York: Springer-Verlag; 2002.
[37] Johnson Pierr. With the public clouds of Amazon, Microsoft and Google, big data is the proverbial big deal. < https://www.forbes.com/sites/johnsonpierr/2017/06/15/with-the-public-clouds-of-amazon-microsoft-and-google-big-data-is-the-proverbial-big-deal > .
[38] Pierre G. Wiki data repository [Online]. < http://www.wikibench.eu/?page_id=60 > .
[39] Google. Data center efficiency. < https://www.google.com/about/datacenters/efficiency/internal/ > .
[40] Facebook. Designing a very efficient data center. < https://www.facebook.com/notes/facebook-engineering/designing-a-very-efficient-data-center/10150148003778920/ > .
[41] Urdaneta G, Pierre G, van Steen M. Wikipedia workload analysis for decentralized hosting. Comput Networks 2009;53(11):1830–45.
[42] Caruana R, Lawrence S, Giles L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. Proceedings of the 13th international conference on neural information processing systems. 2000. p. 381–7.