Yangyang Fu, Xu Han, Kyri Baker & Wangda Zuo (2020)
Assessments of Data Centers for Provision of Frequency
Regulation, DOI: 10.1016/j.apenergy.2020.115621

Assessments of Data Centers for Provision of Frequency Regulation

Yangyang Fu[a], Xu Han[a], Kyri Baker[a,b], Wangda Zuo[a,b,*]

[a] *Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, USA*
[b] *National Renewable Energy Laboratory, Golden, USA*

Abstract

There are numerous opportunities for data centers to participate in demand response programs considering their large energy capacities, flexible working environments and workloads, redundant design and operation, etc. As a type of demand response, frequency regulation requires fast response, and its potential is not fully explored by data centers yet. This paper proposes a synergistic control strategy for data center frequency regulation which uses both IT and cooling systems. It combines power management techniques at the server level with control of the chilled water supply temperature to track the regulation signal from the electrical market. A frequency regulation flexibility factor is also proposed to increase the IT capacity for frequency regulation. The performance of the control strategy is studied through numerical simulations using an equation-based object-oriented Modelica platform designed for data centers. Simulation results show that with well-tuned control parameters, data centers can provide frequency regulation service in both regulation up and down. The performance of data centers in providing frequency regulation service is largely influenced by the regulation capacity bid, frequency regulation flexibility factor, workload condition, and cooling mode of the cooling system, and not significantly influenced by the time constant of chillers. In addition, compared with a server-only control strategy, the proposed synergistic control strategy can provide an extra

[*] Corresponding author
Email address: Wangda.Zuo@Colorado.edu (Wangda Zuo)

regulation capacity of 3% of the design power when chillers are activated. When chillers are deactivated, both strategies have a similar regulation capacity.

Keywords: frequency regulation, data center, frequency regulation flexibility factor, regulation capacity

1. Introduction

The concept of using data centers to provide demand response (DR) services stems from two critical challenges power grids are facing now. First, electric power grids need to balance supply with increasing demand, partially due to increased data center use. Second, the increasing penetration of renewable energy generation in the grid has introduced more fluctuations in the power supply and thus further challenges the power grid management, especially as large-scale energy storage is not readily available.

1.1. Opportunities

Data centers are well-suited candidates to address these two grid-level challenges. The potential for data centers to provide DR encompasses several aspects:

Capacity. Data centers represent very large loads for the grid. In 2010, data centers consumed about 1.1% to 1.5% of the total worldwide electricity and the number was about 1.7% to 2.2% for the U.S. [1]. The design load of an individual data center can be up to 50 MW or more [2]. Further, researches have shown that an optimized 30 MW data center is comparable to 7 MWh large-scale storage in providing DR service for the power grid [3]. One would potentially lose a huge storage capacity for power grid if data centers' large potential capacity for DR is not utilized.

Flexibility. Data centers can be considered as extremely flexible power loads for power grid. They can operate under a broad range of temperatures, which will result in a large range of power load. For example, American Society of Heating, Refrigerating, Air-Conditioning Engineers (ASHRAE) categorizes data centers into four types (A1-A4) based on their requirements of thermal environment. A Class A1 data center typically provides mission critical operations and requires tightly controlled thermal environment. ASHRAE suggests that the allowable supply air temperature in a Class A1 data centers should be within the range of 15 °C to 32 °C [4]. In addition, some data centers have delay-tolerant workloads, which can be shifted in time in response to electricity prices or other grid requests. The delay-tolerant workload is managed by the designs of novel hardware and algorithms that can adapt energy usage in proportion to the utilization of the computing system. Such designs include speed-scaling [5], power-capping [6, 7, 8], moving servers into and out of power saving mode [9], etc. Further, many internet-scale systems that depend on a number of geographically distributed data centers have geographical flexibility to distribute the workload to data centers at different locations [10, 11, 12].

Redundancy. Data centers are designed to meet reliability standards to guarantee their uptime and performance [13]. Most data centers fall into the two high-availability classes defined by the Uptime Institute: Tier III (99.982% availability) and Tier IV (99.995% availability) [14]. Tier specifications address the number and nature of power and cooling distribution, required redundant components, and the ability to repair faults without interrupting IT load. Typical redundant equipment include power sources (e.g., backup generators, Uninterruptible Power Supply (UPS)), power delivery systems (e.g., transformers), chillers, pumps, Computer Room Air Handler units (CRAHs), etc.

Automation. Nearly all sizable data centers (>1 MW IT load) have an Energy Management Control System (EMCS) that monitors and controls the cooling, electrical, and lighting systems [14]. The EMCS system often can provide limited flexibility in system operations to provide other services (e.g. DR).

1.2. Current Status and Gap

Recently, awareness of these potential services has drawn attention to the capabilities of data centers to participate in DR programs. A survey conducted by the Lawrence Livermore National Laboratory in 2015 shows that about 50% of the participating data centers have interest in smart pricing demand side programs, such as load shedding to avoid peak demand [15]. However, data centers are reluctant to participate in incentive-based programs such as providing frequency regulation (FR) in ancillary service market, for multiple reasons.

One reported concern is that data centers are still learning the process of providing FR and that providing grid services on such a fast timescale can be "outside of their visibility or control" [15]. This concern is well-founded considering that these programs provide novel and relatively unexplored territory from the point of view of traditional data center control and operations.

Currently, both academia and industries have limited exploration of FR in data centers as illustrated by our literature review detailed in Section 2. For the cooling system, although the provision of FR using cooling systems in commercial buildings is well-studied, their conclusions might not be applicable to data centers, because of the unique features in data center cooling system, such as equipment redundancy and large internal heat gains. For the IT system, there are still research rooms to improve existing strategies to provide more power flexibility. What's more, there are barely researches about how to provide FR utilizing cooling system and IT system jointly. Towards this, this paper aims to enrich the current literature by introducing a synergistic control strategy for data centers to provide FR service and identifying the influential factors to FR provision.

1.3. Paper Structure

The remainder of this paper is structured as follows. We first provide a detailed literature review on enabling demand side resources in data centers, as well as current state-of-the-art practices regarding data center FR in both academia and industry in Section 2. To track the regulation signal, we develop a synergistic control strategy by adjusting the frequency of the servers and the chilled water supply temperature (CHWST) setpoint simultaneously. Section 4 investigates the performance of a data center participating in a specific regulation market as a new resource. Section 5 compares available regulation capacity of the proposed synergistic control strategy and a server-only control strategy. Lastly, the paper is concluded in Section 6.

2. Literature Review

2.1. Cooling and Electrical System

Data centers are required to operate continuously. The schematic drawing of a typical data center cooling and electrical system is shown in Figure 1 [16]. The cooling system aims to reject a huge amount of heat generated by the servers to the outdoor environment, and the electrical distribution system is designed to power the IT equipment in a safe and reliable manner. The fluid flow in the cooling system is denoted by solid lines, and the power flow in the electrical system is denoted by dashed lines.

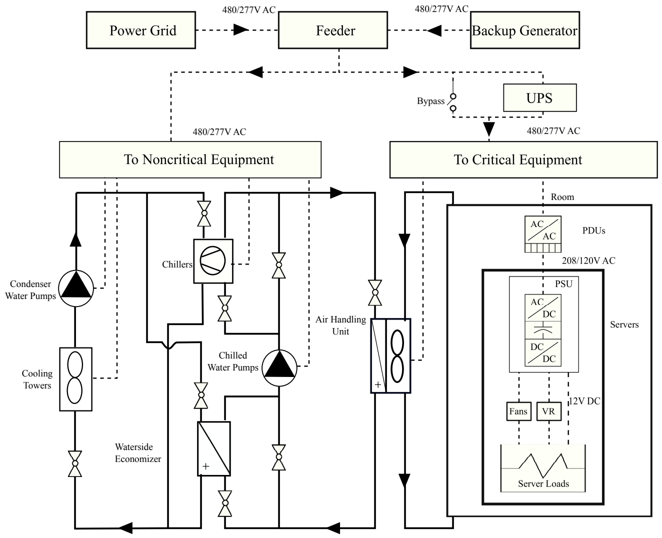


Figure 1: Schematic drawing of the cooling and electrical system in a data center.

Data centers are typically cooled by primary-only chilled water systems with multiple chillers. Some data centers have integrated waterside economizers (WSEs) on the load side. The WSE is located upstream of the chiller on the load side of the common leg as shown in Figure 1. This configuration can allow the WSE to handle the warmest return chilled water and maximize its operational hours. The chiller plant with integrated WSEs can operate in three modes: Free Cooling (FC) mode when only the WSE is enabled for cooling, Partial Mechanical Cooling (PMC) mode when the chiller and WSE are both triggered, and Full Mechanical Cooling (FMC) mode when only the chiller is activated. There are typically redundant pumps/fans to increase the reliability of the fluid delivery system.

The data center is connected to the utility service and the backup generators at the feeder. The incoming power is usually delivered to the data center by a three-phase 480/277V AC system. During normal operation, the UPS is bypassed. The power drawn by noncritical equipment enables the cold chilled water to be produced and delivered to the cooling coils in the CRAHs. The supply air fans are critical equipment and powered to enable the heat transfer between the hot room air and the cold chilled water, and thus deliver the cold air to the servers. In emergency operation, before the backup generators are brought online, the UPS is only utilized to serve the critical equipment (supply air fans and IT equipment), and no power is delivered to the noncritical equipment.

The majority of consumption in the data center is from IT equipment ($\sim$ 50%), then the cooling system ($\sim$ 35%) [14]. The power delivery system has a typical loss of around 11%, and lighting contributes only 4% of the total usage. It is worth to mention that server energy use can fluctuate to some degree as the workload varies. However, most servers and other hardware consume about 60% of their full energy even without workload, which limits the degree of variability. Many energy efficiency techniques are developed to address this issue as detailed in the following sections.

2.2. Demand Side Resources

Ghatikar, Ganti, Matson and Piette [14] summarized some potential demand side resources (DSRs) for data centers to participate in load-shedding and load-shifting. This paper adds additional resources that are capable of providing fast DR, shown in Table 1.

Table 1: Demand side resources in a data center

Target	Strategy	Response Time	Ref
Chillers/CRACs	Adjust compressor frequency/speed	30 s $\sim$ 1 min	[17, 18]
	Reset chilled water supply temperature	$\sim$ 15 min	[19, 20]
	Activate/deactivate redundant chillers/CRACs	2 min $\sim$ 8 min	[21]
Supply air fan	Adjust frequency/speed	2 s $\sim$ 10 min	[22, 23, 24, 25, 26, 27, 28]
	Reset static pressure or flow rate	2 s $\sim$ 10 min	[29, 30, 25, 28]
	Activate/deactivate redundant CRAHs	N/A	N/A
Whole cooling system	Adjust supply/zone air temperature	5 min $\sim$ 50 min	[28, 31]
	Expand/shorten economization hours	5 min $\sim$ 20 min	[14]
UPS	Charge/discharge power	$\sim$ ms	N/A
	Utilize bypass techniques	N/A	[14]
Backup generator	Activate/deactivate backup generators	5 min $\sim$ 15 min	[4, 32]
Transformer	Shut down redundant transformers	N/A	[14]
IT equipment	Apply fine-grained power management	$\sim$ s	[33, 34, 15, 35]
	Apply coarse-grained power management	$\sim$ s	[7, 15, 36]
	Utilize virtualization techniques	2 min $\sim$ 8 min	[21]
	Shut down servers/storage	2 min $\sim$ 8 min	[21]
	Schedule workloads	7 min $\sim$ 22 min	[21]
	Migrate workloads	2 min $\sim$ 175 min	[21]

Site infrastructure, e.g., cooling systems, contribute to a variety of DSRs, in commercial buildings as well as in data centers, including chillers/Computer Room Air Conditioner units (CRACs) [17], CRAHs/fans [23], temperature setpoints at equipment and system level [19], and more [22].

Support loads, such as UPS and power delivery system, are unique resources for providing DR. Using a UPS as an on-site energy storage system can be ideal for DSR because of its capabilities to perform fast charging and discharging. Back-up generators powered by diesel or natural gas are usually configured to start in two to four seconds after a utility outage or voltage fluctuation or for greater than a 10% swing in voltage or frequency [14]. The traditional backup generators at data centers may not be environmentally friendly, in some cases even not meeting Environmental Protection Agency emissions standards [4], which makes this form of response far from ideal. To fully utilize backup generators for in DR programs, it is necessary to reduce their emissions. In addition, a report from Lawrence Berkeley National Laboratory [14] mentions that redundant transformers could be powered down during a DR event to curtail their losses for shedding.

IT infrastructure, such as servers, storage, network etc., provide significant potential for data centers to participate in DR programs. For example, Servers are usually equipped with programmable power management mechanisms, and are capable of adjusting their power consumption using commands from certain interfaces. Fine-grained power management such as Dynamic Voltage/Frequency Scaling (DVFS) at the node level allows the processor to use a lower voltage at the cost of a slower clock frequency by offering high-resolution control [37, 38]. Coarse-grained power management such as power capping at a low resolution and at a more aggregate level can limit the amount of electricity that servers can consume at any given time. Virtualization technologies consolidate and optimize servers, storage, and network devices in real time, reducing energy use by enabling the optimal use of existing data center equipment as shown in [21]. Shutdown of servers and storage or job scheduling are also capable of providing load shedding in response to a DR event, usually by integrating with virtualization technologies. Load migration refers to temporarily shifting workloads from a system on one site to a system on another site. Migration between homogeneous platforms that have the same clusters requires less response time than that between heterogeneous platforms with different clusters.

2.3. Frequency Regulation

FR is a service designed to maintain the frequency throughout the power grid system close to its nominal value (e.g. in the United States, this is 60 Hz). This is achieved by constantly and automatically balancing small fluctuations in supply and demand in real time. The service can be offered by FR resources such as generators on the supply side (which has traditionally been the case) or more recently, by DSRs on the demand side. Providing FR means FR resources are willing to increase or decrease their output (generation for generators, and consumption for DSRs) by following a control signal generated by the market operator.

Different markets operators adopt different policies in FR. This study uses Pennsylvania-New Jersey-Maryland territory, known as PJM. The remaining section introduces a few important and relevant features using PJM as an example. Details of PJM FR service can be found in [39]. PJM divides FR resources into two categories: ramp-limited and capacity-limited. Ramp-limited resources respond slowly to FR signals but with a large capacity. Examples are coal-fired steam power plants. Capacity-limited resources, including batteries, flywheels, and responsive loads, have small capacities but can respond to FR signals in a quick manner. PJM has developed two types of FR signals for these two resources: traditional regulation A signal (RegA) for ramp-limited resources and dynamic regulation D signal (RegD) for capacity-limited resources. Under these two FR signals, ramp-limited resources mostly get paid for their capacity, and capacity-limited resource mostly get paid for their performance.

In the PJM market, new resources aiming to enter the regulation market need to have a capacity of at least 100 kW and to pass an initial test by obtaining at least 0.75 for a defined performance score [39]. The initial test signals of

RegA and RegD are available at [40]. The performance score is calculated as a composite score of accuracy, delay and precision, which are shown below [39].

$$c_{sig,res} = \frac{COV(reg, res)}{\sigma_{reg}\sigma_{res}} \tag{1}$$

$$S_{accuracy} = \max_{\delta=0-5\text{ min}} (c_{reg,res(\delta)}) \tag{2}$$

$$S_{delay} = \left| \frac{5\text{ min} - \delta^{*}}{5\text{ min}} \right| \tag{3}$$

$$S_{precision} = 1 - \frac{1}{n} \sum \left| \frac{res - reg}{\overline{reg}} \right| \tag{4}$$

$$S_{performance} = \frac{S_{accuracy} + S_{delay} + S_{precision}}{3} \tag{5}$$

In the above equations, reg represents the regulation signal the DSRs receive from the electrical markets, and res represents the response signal the DSRs generate after control actions. c, COV and σ are the correlation coefficient, covariance, standard deviation of these two signals. In PJM, the response signal res is recalculated with a time shift δ ranging from 0 to 5 minutes in an increment of 10 seconds, which leads to 31 response signals $res(\delta)$. The accuracy score $S_{accuracy}$ is the maximum correlation coefficient c between reg and $res(\delta)$. The delay score S_{delay} is calculated based on the delay time δ^{*} when the maximum accuracy score is obtained using Eq. (3). The precision score $S_{precision}$ is defined as the relative difference between regulation signal and response signal, where n is the number of samples in the hour, and $\overline{reg}$ is the hourly average regulation signal. The final performance score $S_{performance}$ in that hour is calculated as the weighted average of the three individual scores.

2.4. Frequency Regulation in Data Centers

Data centers have a rich pool of DSRs as shown in Table 1. FR service that requires fast responses from data centers can be individually or jointly provided by site infrastructure, support loads, and IT infrastructure. This section summarizes the state-of-art research of using data centers to provide FR.

2.4.1. Site Infrastructure

There are few studies in using data center site infrastructure (mainly cooling systems) to provide FR. However, the data center cooling systems are also commonly used in commercial buildings. Providing FR with commercial building cooling system has been well studied and the knowledge can be potentially applied to the data centers. Thus, this subsection mainly discusses the efforts of providing FR service by the cooling systems in commercial buildings.

Zhao, Henze, Plamp and Cushing [31] proposed two methods of using HVAC systems in commercial buildings to provide FR: direct methods, such as adjusting static pressure setpoint, and indirect methods, such as adjusting zone air temperature setpoint. Many experimental studies focused on using the supply air fan to provide FR by changing static pressure in the air duct [28], air flow rate setpoint [25] or frequency of the motor [26]. The response time can be as low as 2 s [22] by directly adjusting the VFD frequency. Su and Norford [19, 20] designed and evaluated a FR controller to adjust the CHWST setpoint for a chiller to track a FR signal.

Due to the specific nature of data centers, there are a few challenges and opportunities in adopting the outcome of the studies in commercial building to data centers. Data centers are required to have sufficient redundant capacity of the cooling equipment (CRAHs/CRACs, pumps, and even chillers) to satisfy reliability requirements, which is not necessary for commercial buildings. Additionally, the commonly-used control strategies in commercial buildings and residential buildings leverage to some extent the passive thermal mass in the room (e.g., building envelope) to mitigate the effects of control interruptions on the thermal environment. This may not be applicable to data centers directly, because the large internal thermal heat gains in the data center room can neglect the thermal delay impact of its passive thermal mass.

In addition, most studies [19, 20] performed the evaluation of the FR service only on isolated equipment, such as chillers, and their energy influence on the overall cooling system were separated. For instance, when regulation down is required, the chillers can raise their supply temperature setpoints to decrease power consumption. However, the supply air fans have to increase their power as a response to the increased chilled water temperature in the cooling coils, which counteracts the efforts of reducing the system power consumption. Thus, it is difficult to quantify the net benefits from electrical markets without considering the system as a whole.

2.4.2. IT Infrastructure

There are several techniques available to adjust the server power in order to limit data center power usage as mentioned in Table 1. Some of the techniques (e.g. DVFS and dummy workload) can also be used to provide FR service because of their fast response.

Data center IT infrastructure is not always running at its full capacity. It is possible to use that unutilized capacity to provide FR service. A few studies focused on the FR service by using power management techniques such as DVFS [34, 41]. DVFS is widely used to dynamically adjust server power consumption with required performance. Since power consumption of a processor varies quadratically with voltage/frequency but gate delay varies only linearly, the processor's voltage/frequency can be adjusted to change the power while maintaining the adequate performance for the current workload [42]. One recent publication investigated the possibility of providing FR service by introducing extra dummy workload to the servers [35]. The purpose of the dummy loads is to adjust the server utilization rates so that the server power can respond to external signals. However, there is still more spaces in using this resource to provide FR service which will be discussed in this paper.

2.4.3. Support Loads

Traditional usage of UPS in data centers is to serve as backup power. However, UPS can also be potentially used for peak shaving, power regulation, and assisting with renewable integration [43]. Studies showed that one could reduce operational costs by up to 30% by using UPS for peak demand shaving and FR service together [44]. An architecture for distributed per-server UPS is presented in [45], which can participate in ancillary service markets without degrading the Quality of Service (QoS).

When considering degradation of the equipment, Chen, Liu, Coskun and Wierman [46] concluded that it may not be economical for batteries to participate in some markets. Narayanan, Wang, Mamun, Sivasubramaniam, Fathy and James [43] suggested that flow battery, a new type of electrochemical cell, allows for a fast response and placement flexibility, while conventional electrochemical energy storage technologies, such as Lead-Acid and Lithium-Ion, used for power backup, are less suitable for FR service.

There are also significant efforts underway within industry to improve the UPS design in order to enable FR service. In 2017, power management specialist Eaton launched the first pilot project of "UPS-as-a-Reserve" service [47], an initiative that enables data center owners to participate in regulation of the power grid while getting paid for their contribution. New battery technologies such as the Tesla Powerpack can charge or discharge instantly to provide FR service, voltage control, and provide spinning reserve services to the grid due to its low ramp time. However, such systems currently require large and expensive batteries to offer significant regulation capacity to the market.

2.4.4. Synergistic Strategies

Only a few of papers studied synergistic strategies that combine IT infrastructure and support loads for FR service. For example, Guruprasa, Murali, Krishnaswamy and Kalyanaraman [48] developed a coupled data center and battery system, which allows data center to work in conjunction with a small battery to provide fast FR service. Li, Brocanelli, Zhang and Wang [49, 50] also considered the joint power management of a data center and plug-in electric vehicles for FR service.

It is worth to mention that the strategy that combines the IT infrastructure and site infrastructure (e.g. cooling systems) are not well studied yet because of several concerns towards manipulating the cooling system in data centers. On top of that might be the concern of cooling safety. For example, adjusting the room temperature might introduce hot spots in the racks. In addition, for an energy efficient data center whose power usage effectiveness is small, the power of the cooling system is relatively small compared with IT equipment. However, totally ignoring the capability of the cooling system might be a waste of existing resources since data center cooling energy still accounts for about 35 $\sim$ 40 % of the data center overall energy usage in the worldwide [51]. In this paper, a synergistic control strategy that combines the operation of the cooling system and IT equipment is proposed and the extra benefits of including the cooling system in the regulation control is also studied.

3. Proposed Synergistic Control Strategy

In this section, we propose a synergistic control strategy for data centers to provide FR service, which is evaluated at a whole system level in the Section 4. This strategy is composed of four major parts. The first one is *Baseline Routine*, which predicts the baseline power usage when the data center provides no FR. The second one is *Bidding Capacity*, which is the capacity bid that the data center submits to the electrical market. The third one is *Server Power Management*, where an aggregator is adopted to represent the aggregated performance of servers in the data center. The clock frequency of the aggregator can be directly changed by a Proportional-Integral-Derivative (PID) controller in order to follow the regulation signal. Based on that, the desired frequencies for individual servers will be determined by a set of predefined assignment rules and then be propagated to all servers using techniques such as DVFS. The forth one is *Cooling Power Management*, which adjusts the CHWST setpoint to respond to the regulation signal.

Figure 2 shows the workflow of the proposed synergistic control strategy. The *Baseline Routine* outputs the prediction of the overall power profile for the data center P_{bas} when no FR service is provided. In this paper, the prediction is performed using detailed energy models, although many other methods such as machine learning techniques can also be used. The detailed energy models and baseline settings can be referred to Section 4.1. The *Bidding Capacity* is a module that can calculate the optimal capacity bid for the data center at each time step, and output raw regulation power $\Delta P_{reg,Raw}$ based on the optimal capacity bid and received regulation signal r from the electrical market. In this paper, we assume the capacity bid C_{reg} is known, since finding the optimal bid is not the focus here. Then, the reference power P_{ref} for the data center to track is the summation of the predicted baseline power P_{bas} together with the raw regulation power $\Delta P_{reg,Raw}$.

The *Server Power Management* first determines the number of required active servers in the aggregator N_{act} based on the predicted workload $\lambda^{'}$ in the next time step (e.g., one hour ahead). Then a closed-loop control using a PID controller is utilized to minimize the error between the measured total power usage P_{mea} and the reference power P_{ref} by adjusting the aggregated frequency of the server aggregator. Meanwhile, the *Cooling Power Management* applies an open-loop control to adjust the cooling system power usage by resetting the CHWST setpoint in response to the received regulation signal r.

The server aggregator receives the aggregated frequency f_{agg} and the required number of active servers N_{act} from the FR controller. Assuming there are N_0 number of servers in the data center, the server aggregator then calculates the CPU frequency f_i for an individual server i based on predefined assignment rules. The cooling system receives CHWST setpoint from the FR controller. Both the IT system and the cooling system respond in such a way that their total power P_{mea} is adjusted to track the reference power P_{ref}.

For the aggregator, there are several assignment rules to control the individual server's frequency [34, 35]. We can also represent the aggregated server power $P_{servers}$ of all servers under an assignment rule using a simplified model [34] and this approach is adopted by this paper and detailed in Section 9.1.1. For the FR controller, more details are described in the rest of this section.

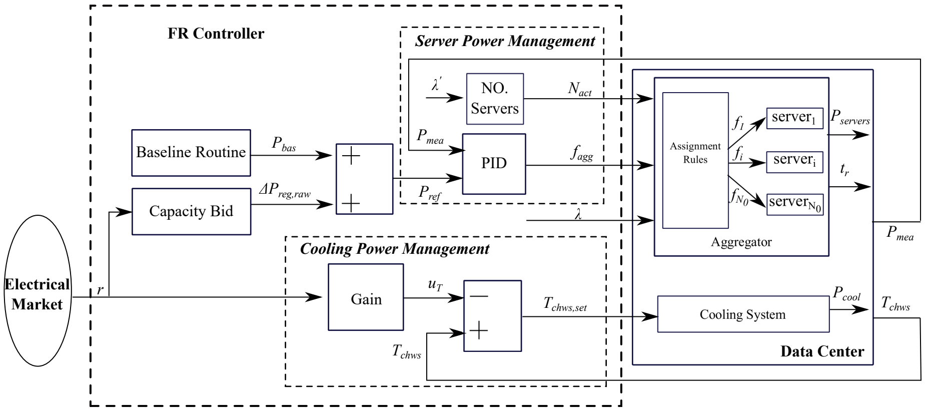


Figure 2: Data center frequency regulation control

3.1. Server Power Management

The servers in the data center are represented by an aggregator, which is characterized by the active number of servers N_{act}, and the aggregated frequency f_{agg} as shown in Section 9.1.1. Base on these two parameters, the aggregator can output the total power of the servers $P_{servers}$ and the average service response time t_r. The *Server Power Management* is used to determine N_{act} and f_{agg} at each time step based on the normalized raw regulation signal received from the electrical market, r, ranging from -1 to 1, and incoming actual workload λ.

3.1.1. Reference Power

The reference power P_{reg} is calculated as

$$\Delta P_{reg,raw}(t) = r(t)C_{reg} \tag{6}$$

$$P_{ref}(t) = P_{bas}(t) + \Delta P_{reg,raw}(t) \tag{7}$$

where $\Delta P_{reg,raw}$ is the raw power signal and C_{reg} is the regulation capacity that the data center bids in the market.

3.1.2. Number of Active Servers

The number of servers in a data center needs to satisfy the following condition in order to ensure the stability of the IT service. This condition means that the service capability $N_{act}(t)\mu(t)$ in the data center should be greater than the workload) $\lambda(t)$:

$$N_{act}(t)\mu(t) > \lambda(t), \tag{8}$$

where $\mu(t)$ is the actual service rate, which denotes the number of requests that a single server can process every second. The service rate is typically proportional to the server's CPU frequency, as defined in Eq. (23) [34, 35].

Under design conditions, to guarantee reliability, a scaling factor γ as defined in Eq. (9) is utilized here to describe the design redundancy of the servers [34]. The γ is set to greater than 1. If $\gamma = 1$, it means all the CPU clock frequencies need to set at the maximum level just to serve the average workload, which limits the potential of FR. The γ is defined as

$$\gamma = \frac{\mu_0 N_0}{\lambda_0}, \tag{9}$$

where μ_0 is the nominal service rate of a single server, N_0 is the nominal number of servers in a data center room, and λ_0 is the nominal workload to be served by the data center.

When using a server aggregator model as described in Section 9.1.1, the γ can then be rewritten as:

$$\gamma = \frac{kN_0}{\lambda_0} = \frac{kN_{act}(t)}{\lambda^{'}(t)}, \tag{10}$$

where k is a constant parameter, assuming the service rate is proportional to the aggregated frequency, N_{act} is the number of active servers at current time step, and $\lambda^{'}$ is the predicted workload, here we use the mean workload of the current time step as the prediction.

The number of active servers is calculated at an interval of 1 hour because the servers have relatively long wakeup time. The detailed formula is shown in Eq. (11):

$$N_{act}(t) = \lceil \frac{\gamma\lambda^{'}(t)}{k} \rceil, \tag{11}$$

where the operator $\lceil x \rceil$ is the ceiling function that gives the least integer greater or equal to x.

However, when aiming to provide FR service, Eq. (11) cannot fully exploit the design redundancy introduced by γ. Here we propose to revise it by adding a FR flexibility factor β during the operation:

$$N_{act}(t) = \lceil \beta \frac{\gamma \lambda^{'}(t)}{k} \rceil, N_{act}(t) \in [0, N_0]. \tag{12}$$

The greater β is, the more servers are activated for a specific workload. The intention of introducing β is to increase the FR capacity of IT servers. Detailed influence of β on the FR service performance will be investigated in Section 4.

3.1.3. Aggregated Frequency Control

A PID controller is used to follow the reference power P_{ref} by directly changing f_{agg} at an interval of 4 s.

$$f_{agg}(t) = K_p e(t) + K_i \int_0^t e(x)dx + K_d \frac{\mathrm{d}e(t)}{\mathrm{d}t}, f_{agg}(t) \in [f_{min}, f_{max}] \tag{13}$$

$$e(t) = P_{ref}(t) - P_{mea}(t) \tag{14}$$

In the above equations, K_p, K_i, and K_d denote the coefficients for the term P, I and D, respectively. e is the control errors between P_{ref} and P_{mea}. The maximum aggregated frequency is 1, while the minimum frequency varies based on the number of active servers due to the constraints of QoS. Details on how to determine f_{min} are described in Section 3.1.4.

3.1.4. Minimum Aggregate Frequency

The same approach in Ref. [34] is used here to find the minimum allowable aggregated frequency. Using a service response time model shown in Eq. (23) and Eq. (28), we know that the response time of the servers depends on the aggregated frequency. If the frequency is low, then it takes relatively long time for the servers to respond to the arrival workload, which means the QoS of the data center is compromised. To enable FR and guarantee the QoS, the aggregated frequency should meet a minimum value. From Eq. (23) and Eq. (8), we can get

$$f_{agg}(t) > \frac{\lambda(t)}{kN_{act}(t)} \tag{15}$$

Combining Eq. (15) and Eq. (12), we can obtain a lower bound for the aggregated frequency:

$$f_{agg}(t) \geq \frac{\lambda(t)}{\beta\gamma\lambda'(t)} \tag{16}$$

To ensure the QoS while providing FR, the response time should satisfy:

$$t_r \leq t_u \tag{17}$$

where t_u is the upper response time limit of the data center.

In Eq. (28), the service time $t_s = \frac{1}{\mu(t)}$ accounts for the majority of the response time [34]. A necessary condition to guarantee the response time constraint is that

$$t_s = \frac{1}{\mu(t)} \leq t_u \tag{18}$$

By substituting Eq. (23) into Eq. (18), we can get another lower limit of the aggregated frequency:

$$f_{agg}(t) > \frac{1}{kt_u} \tag{19}$$

Combining Eq. (16) and (19), we can define f_{min} as:

$$f_{min} = \max(\frac{\lambda(t)}{\beta\gamma\lambda'(t)}, \frac{1}{kt_u}) \tag{20}$$

3.2. Cooling Power Management

The cooling system power is managed by resetting the CHWST. The regulation signal from the electrical market is directly used to change the CHWST setpoint $T_{chws,set}$ by Eq. (21).

$$T_{chws,set}(t) = T_{chws}(t) - r(t)\Delta T, \tag{21}$$

where T_{chws} is the CHWST at current time step, ΔT is the user defined regulation range for the temperature, and varies based on the design supply temperature range of chillers. Here we set $\Delta T = 2$ °C. The negative sign at the right term means when regulation up is needed, the temperature setpoint should be reduced, and vice versa.

4. Case Study

The purpose of this case study is to investigate the performance of the proposed control strategy for tests with RegA and RegD signals. Via the case study, we try to understand how the FR service performance can be affected by some important factors, such as regulation capacity bid, FR flexibility factor, thermal response time of the chiller, workload condition, and cooling mode of the cooling system.

4.1. Simulation Setup

The considered data center is located in Chicago, which is in ASHRAE Climate Zone 5A (Cool Humid) and within the PJM market territory. The configuration of the cooling system is shown in Figure 1. The number of servers in the data center is 8000. The design factor γ is set to 1.5 [34]. The total nominal electrical load is 2680 kW, with a design power usage effectiveness of 1.35. The calibrated coefficients for Eq. (22) are $b_0 = 0.0154$, $b_1 = 1.5837$, $b_2 = 0.1373$, $c_0 = -22.3540$ and $c_1 = 121.0212$ using the method mentioned in Ref. [34], with a mean absolute percentage error of 3.6%. When not providing FR, the server aggregator operates at a frequency of 0.8, and the CHWST setpoint is set to 8 °C. A typical workload from a web service data center is normalized and used here as shown in Figure 3 [34]. The test signal for 2019 is downloaded from the PJM homepage as shown in Figure 4 [40]. To guarantee the QoS of the data center, the maximum average response time is set to 6 ms. The simulation is performed in a previously verified and calibrated Modelica-based environment [52, 53, 16].

This paper adopted the cooling system control in a previous study [54]. The cooling equipment ON/OFF control are characterized by minimum runtime constraints and minimum downtime constraints. When a device is turned ON, it must remain ON for a certain amount of time; similarly, it must remain OFF for a certain amount of time that when it is turned OFF. These constraints can mitigate the potential risks of fast changing FR signals on the wearout of

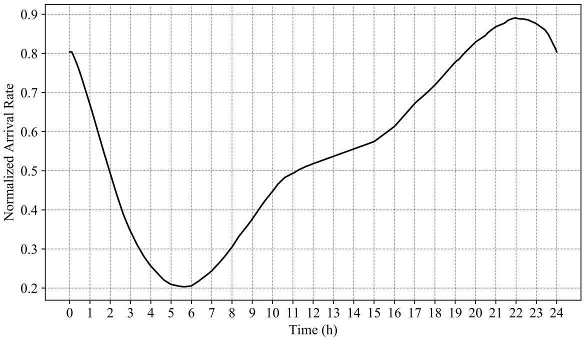
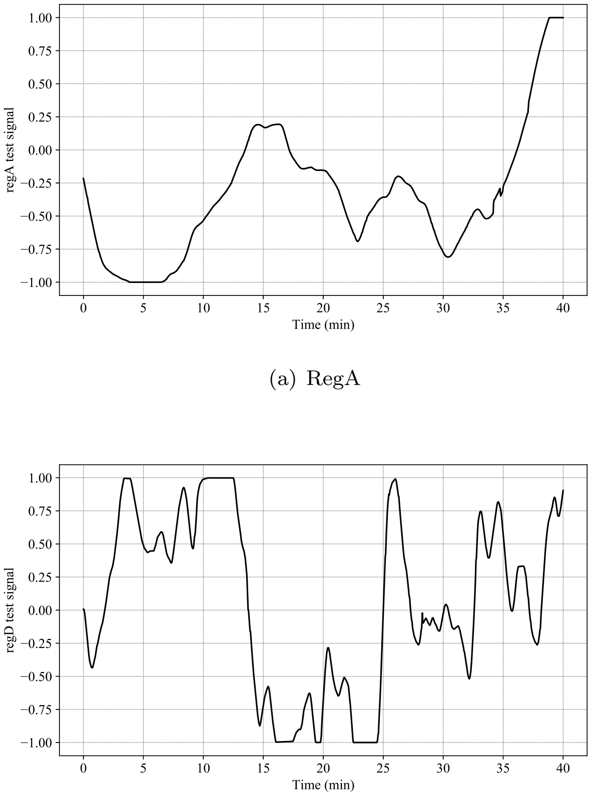


Figure 3: Normalized daily workload



(a) RegA

(b) RegD

Figure 4: Raw test signal from the PJM market

cooling equipment [55, 56, 57]. In this paper, we set the minimum runtime and minimum downtime of major cooling device (e.g., chillers, WSE) to 20 minutes. The cooling tower fan speed is controlled to satisfy the requirement of temperature setpoint under the maximum fan speed. In FC mode, the fan speed is controlled to maintain a predefined CHWST at the downstream of the economizer, and not exceed the predefined maximum fan speed that is 90% of the normal speed. In PMC and FMC modes, the fan speed is controlled to maintain the supply condenser water at its setpoint. The FR performance will be influenced by this local control in the cooling system because the CHWST setpoint is adjusted in the proposed strategy.

4.2. Simulation Scenarios

We swept the following parameters (Table 2) to investigate the FR performance in the studied data center. The regulation capacity is chosen as 5%, 10%, and 15% of the design electrical load, which is 134 kW, 268 kW and 402 kW, respectively. The FR flexibility factor β is set to 0.9, 1.0, and 1.1, to investigate the influence of the operational redundancy on the performance of regulation. The chiller's thermal response time τ is set to 5 min, 10 min and 15 min, which can reflect different types of chillers as referred in [19, 20]. As shown in Figure 3, three different workloads are analyzed and compared: light, medium and heavy, which happen during 5:00-6:00, 12:00-13:00 and 22:00-23:00, respectively. The simulation is conducted for a cold day when the cooling system is in FC mode, and a hot day when the cooling system is in FMC mode. Both RegA and RegD are evaluated in this case study.

4.3. Results and Discussion

Using the above settings, numerical simulations were performed and the results are presented in the following subsections.

4.3.1. Regulation Capacity Bid

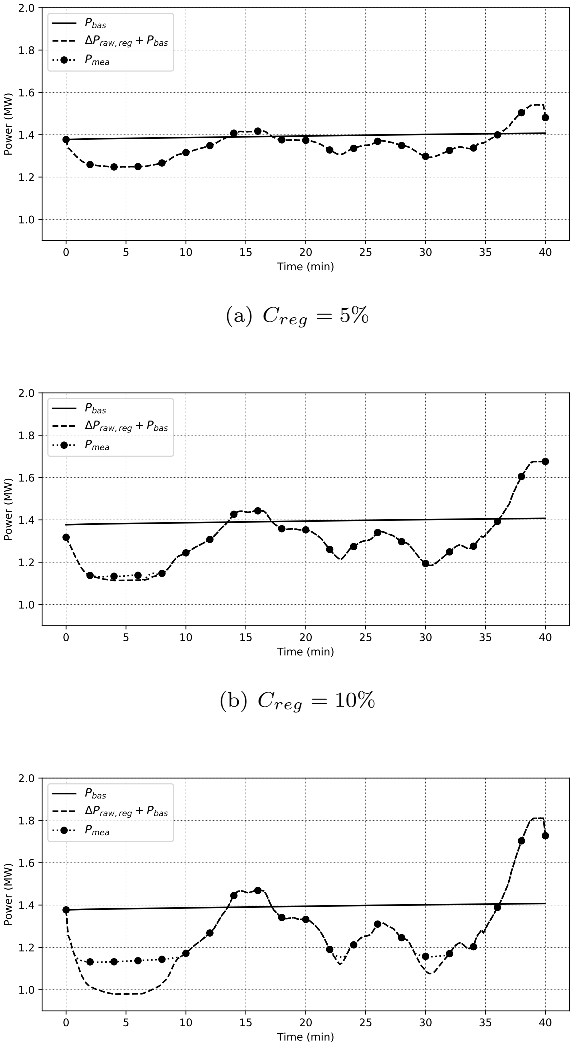
The regulation capacity bid C_{reg} has a major influence on the FR service performance in PJM market, especially when regulation down is required. The

Table 2: Swept parameters for FR service

Parameters	Values	Units	Comments
C_{reg}	134, 268, 402	kW	regulation capacity
β	0.9, 1.0, 1.1	–	FR flexibility factor
τ	5, 10, 15	min	time constant of the chiller's response to the CHWST setpoint
workload	light, medium, heavy	–	requested IT service
cooling mode	FC, FMC	–	cooling mode that determines the activation and deactivation of different cooling sources

larger the bidding regulation capacity is, the worse the service performance is. For example, in Table 7, the performance score defined by Eq. (5) decreases from around 0.98 to around 0.89 as C_{reg} increases from 5% to 15% at $\beta = 1.1$.

The decrease is due to insufficient regulation capacity as shown in Figure 5. When the bidding capacity is 5%, the system can generally track the reference signal in an accurate way. Both regulation down and regulation up requests can be met. When the bidding capacity is 10% and 15%, the regulation performance is mainly influenced by regulation down because the maximum regulation down capacity is achieved at the minimum aggregated frequency required by the constraints of QoS. Because there are sufficient number of active servers, regulation up can be met for both bids (10% and 15%). If the bid is further increased to a relatively large value, with specific amount of active severs operating at their maximum frequency, the data center cannot meet the regulation up either. Figure 5 also shows that the data center provides asymmetric regulation up and regulation down capacity. It is easier to provide regulation up because of the design redundancy introduced by γ. Regulation down is constrained by the QoS.

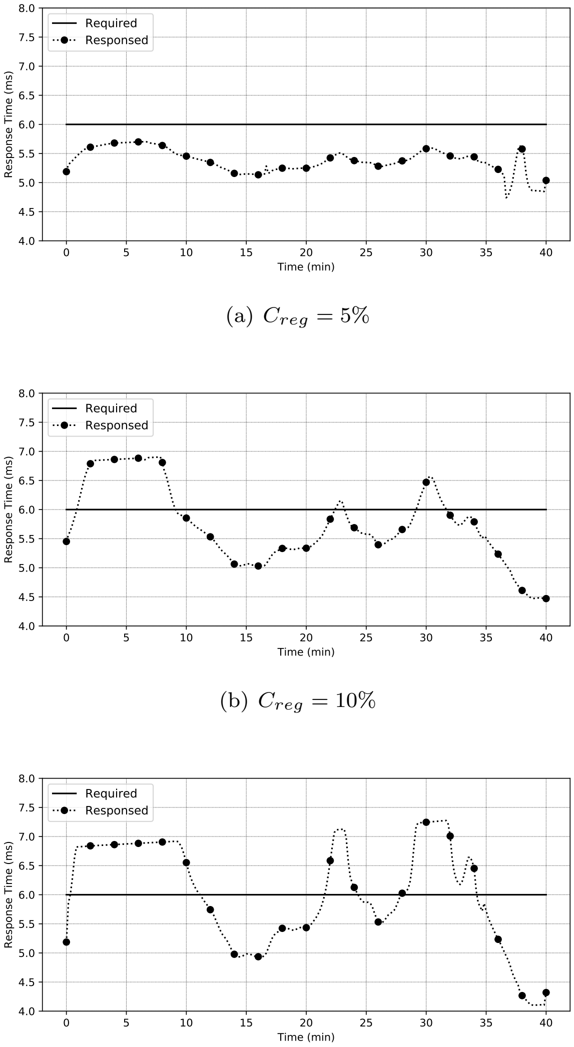


(a) $C_{reg} = 5\%$

(b) $C_{reg} = 10\%$

(c) $C_{reg} = 15\%$

Figure 5: Detailed signal tracking for RegA test at $\beta = 1.1$, $\tau = 5$, medium load, FMC mode and different bids



(a) $C_{reg} = 5\%$

(b) $C_{reg} = 10\%$

(c) $C_{reg} = 15\%$

Figure 6: Detailed response time for RegA test at $\beta = 1.1$, $\tau = 5$, medium load, FMC mode and different bids

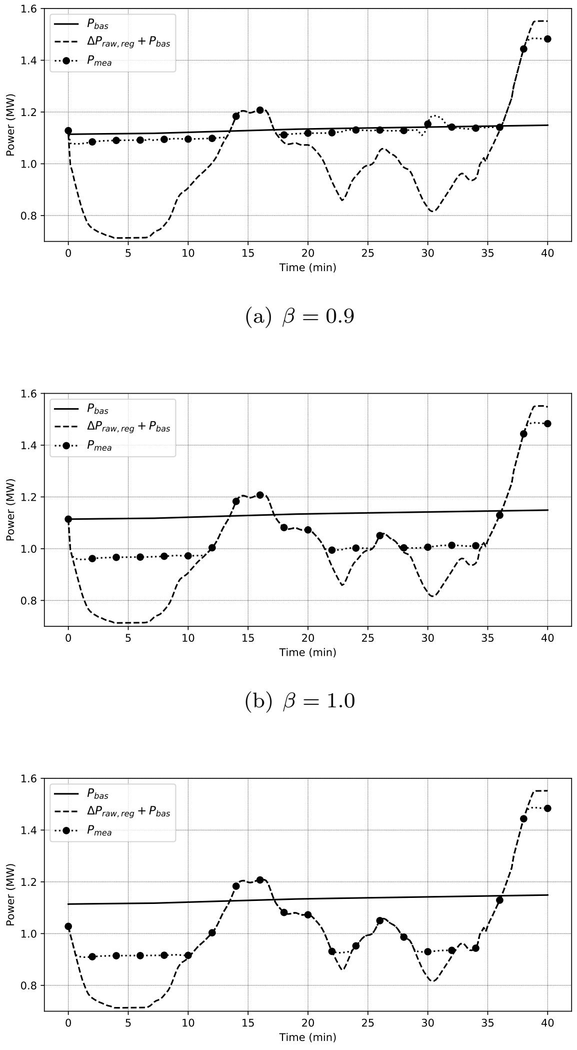
Figure 6 shows the average response time of the servers whiling providing FR service. When the bid is small (e.g., 5%), the average response time is within 6 ms as required. However, when the bid increases, the proposed strategy cannot strictly constraint the average response time to be within 6 ms. The violations happen only when the system cannot provide sufficient regulation down capacity. When regulation down is required, the servers work at a low frequency constrained by the minimum frequency as defined in Eq. (18). However, because Eq. (18) is not a sufficient condition, it happens that the system would violate the constraints when it cannot provide enough regulation down capacity.

4.3.2. Flexibility Factor

The larger β is, the more servers are activated for a specific workload. When more servers are activated, based on Eq. (22), more base power related to the amount of active servers is needed, which can increase the regulation up capacity. In addition, more active servers means a smaller minimum aggregated frequency as shown in Eq. (20), which can decrease the power usage for the frequency-related term as shown in Eq. (22). The tradeoff of power consumption between the decreased frequency and increased number of active servers determines if regulation down capacity can be increased or decreased.

Generally, the proposed β can significantly improve the FR performance for almost all the scenarios when it is appropriately tuned. Figure 7 shows detailed signal tracking performance for a RegA test with a medium load in FC mode. The regulation down capacity can be increased as β increases, which then significantly improves the performance score as shown in Table 4.

For most scenarios, when β is less than 1, there is a significant degradation of the FR service compared to $\beta = 1.0$. For example, in Table 5, the performance score is only around 0.55 when β is 0.9. The reason can be seen in Figure 7(a), which shows that the system can barely provide regulation down service at a small β. This inability is due to the large allowable minimum aggregated frequency for FR service that is illustrated in Figure 8. The minimum aggregate frequency decreases from about 0.80 to 0.67 as β increases from 0.9 to 1.1. When



(a) $\beta = 0.9$

(b) $\beta = 1.0$

(c) $\beta = 1.1$

Figure 7: Detailed signal tracking for RegA test at $C_{reg} = 15\%$, $\tau = 5$, medium load, FC mode and different β

β is 0.9, and regulation down service is required, the server aggregator can only work at its minimum frequency (around 0.80), which leads to a similar power consumption compared with the baseline. When β is 1.1, because the minimum frequency is decreased to 0.67, more regulation down capacity can be provided.

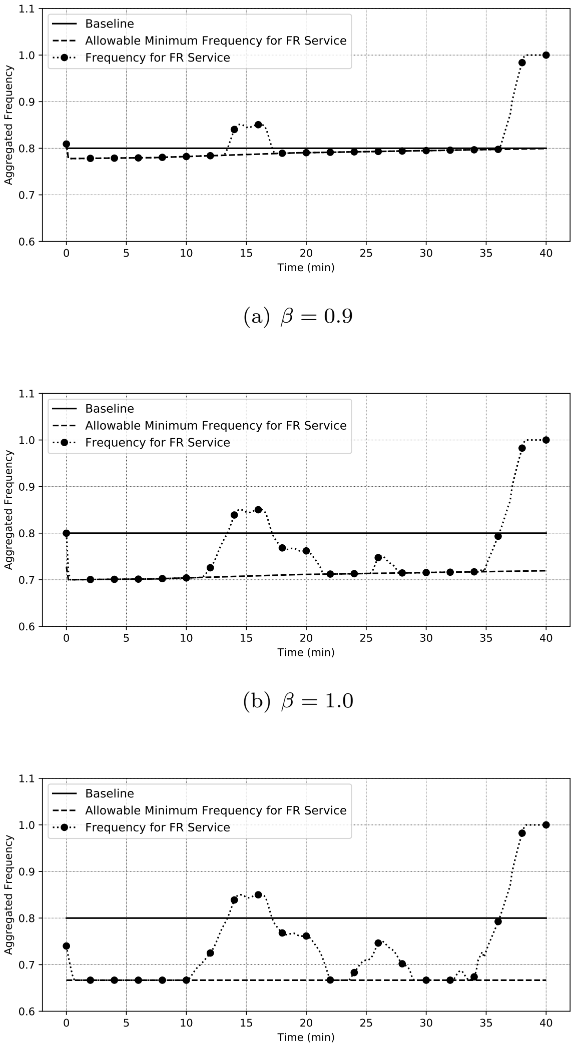
When the bidding regulation capacity C_{reg} is small (e.g. 134 kW) at medium and heavy workload, the performance of FR service is almost the same for β=1.0 and β=1.1 (e.g., Table 4 and Table 5). The reason is when C_{reg} is small, the aggregated frequency only needs be somewhere between the minimum frequency and the maximum frequency, which can track the reference power so well that both regulation down and regulation up can be met.

4.3.3. Other Parameters

Thermal time constant of the chillers: It has little influence on the FR service using the proposed synergistic control strategy as shown in Table 3 to Table 14. In FC mode, the chillers are off, therefore little influence can be observed. In FMC mode, the chillers are activated to provide cooling. Although they are slow-response resources, the delays can be compensated by the fast-response resource, e.g., servers that act like a battery system.

Workload: The larger the workload is, with the same amount of bidding capacity, the better the regulation performance is. For example, Figure 9 compares the detailed power signal tracking for a RegD test under different workloads but with the same C_{reg}, β, τ and cooling mode. With a larger workload, more servers are activated based on Eq. (12), which can subsequently provide more regulation up and regulation down, thus improving the regulation performance.

Cooling mode: The proposed strategy for both RegA and RegD test can provide better performance in FMC mode than FC mode. For example, by comparing Table 7 and Table 4, we can find out that the performance score is about 0.90 in FMC mode and about 0.83 in FC mode when bidding 268 kW with a β of 1.0. The reason is that when the data center works in FMC mode, where chillers are activated to provide cooling, the proposed strategy can provide more regulation down capacity as shown in Figure 10.

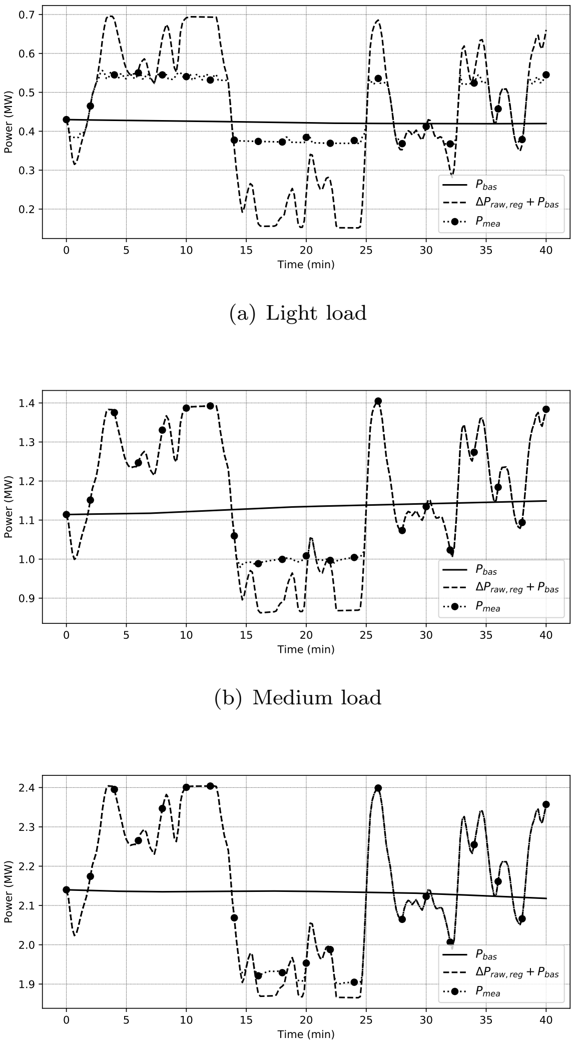


(a) $\beta = 0.9$

(b) $\beta = 1.0$

(c) $\beta = 1.1$

Figure 8: Controlled frequency for RegA test at $C_{reg} = 15\%$, $\tau = 5$, medium load, FC mode and different β

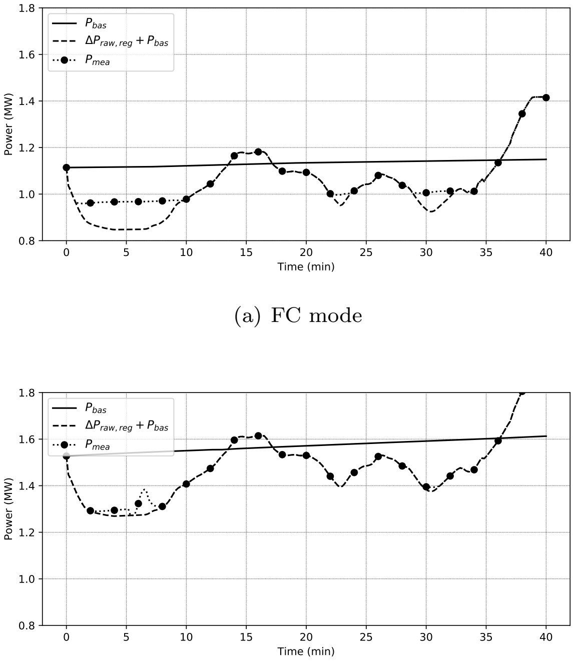


(a) Light load

(b) Medium load

(c) Heavy load

Figure 9: Detailed signal tracking for RegD test at $C_{reg} = 10\%$, $\beta = 1.0$, $\tau = 10$, FC mode, and different workloads



(a) FC mode

(b) FMC mode

Figure 10: Detailed signal tracking for the RegA tests at C_{reg} = 10%, β = 1.0, τ = 10, medium load, and different cooling modes

In summary, the simulation results show that the data center using the proposed control strategy is able to participate in the PJM regulation market as a new resource when the parameters are well-tuned. The regulation performance is largely influenced by the bidding regulation capacity C_{reg}, FR flexibility factor β, workload condition and cooling mode of the cooling system, and minimally influenced by the thermal time constant of chillers τ.

5. Synergistic Control versus Servers-only Control

As stated in the literature review, most of current researches focus on servers-only strategies, where only the CPU frequency of the IT servers is adjusted to respond to the regulation signal. Few studies consider synergistic control strategies, especially the cooling system and the servers, for FR service in data centers. This section aims to numerically investigate the benefits of including cooling system in the FR control strategy. Two strategies are compared in this section. s_1 - the servers-only control strategy that only adjusts the aggregator frequency to respond to regulation signal; s_2 - the proposed synergistic control strategy that adjusts the aggregator frequency and reset CHWST simultaneously to respond to regulation signal. The strategy that only resets CHWST is not considered here, because of its relatively slow response and small capacity. The comparison of s_1 and s_2 focuses on the different maximum regulation capacities for each strategy under FMC mode and FC mode.

The regulation capacity is identified as the maximum symmetric power range that the data center can operate within. Data centers have a nonlinear baseline since cooling system operation changes intra-hour and hour-to-hour in response to varying weather, equipment staging and workloads. Similarly, potential regulation capacity varies throughout the day as a function, for example, of weather, workloads, and how the cooling system respond to the workloads and weather etc. In this section, we use a simulation-based environment to determine the regulation capacity for each time step. Here we set the time step to 1 h, because typically in PJM regulation market, the regulation capacity bid can be updated on an hour basis.

The regulation capacity can be found using a model perturbation method as introduced in [58]. The model perturbation method uses mathematical models to study the relationship between the data center system power response and changes in the control inputs (e.g., f_{agg} in s_1) at each time step. Figure 11 illustrates the process of determining the regulation capacity of the data center using s_1. For example, at 12:00, the model is assumed to be tracking a baseline aggregated frequency, and it intends to determine the regulation capacity for the current hour (12:00-13:00). The control input f_{agg} is adjusted from the baseline by simulating 0.02 increments between a minimum aggregated frequency (calculated from Eq. (20) to a maximum aggregated frequency of 1. Simulation power responses are then compared with the baseline power to determine the regulation up capacity and regulation down capacity. The symmetric regulation capacity of the whole system is then determined by the minimum of the regulation up and regulation down capacity. The same approach is also used for s_2, which adjusts an additional control input CHWST with an increment of 0.5 °C between an lower limit of 6 °C and 10 °C.

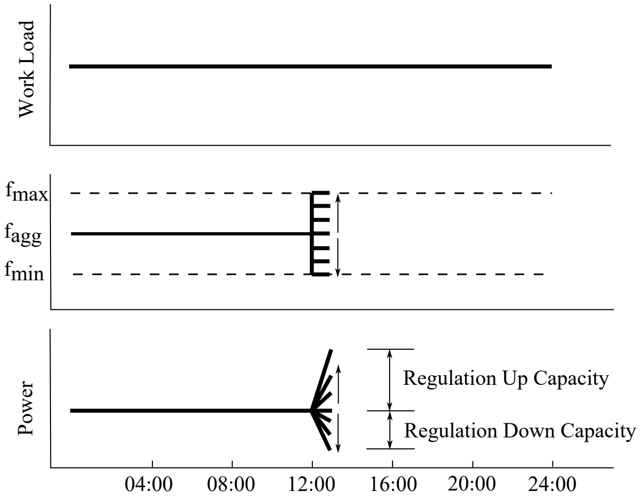
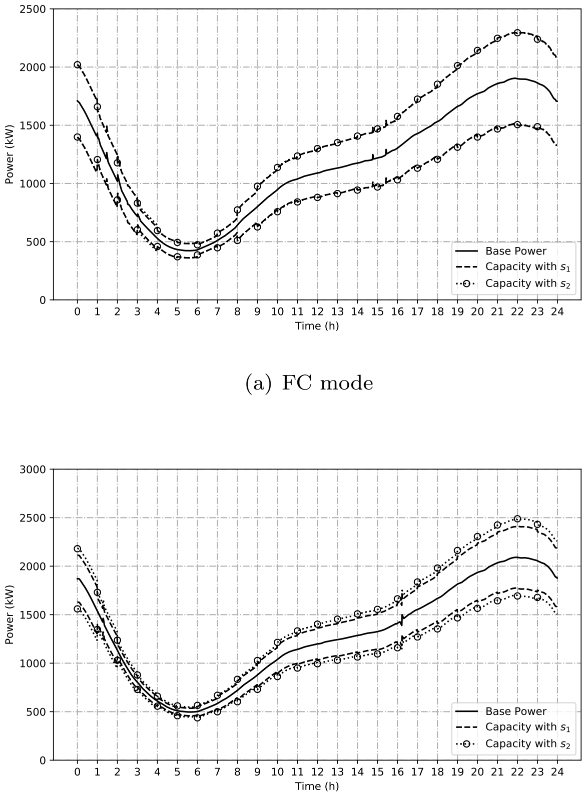


Figure 11: Illustration of model perturbation method to find the regulation capacity

Figure 12 shows the regulation capacity for the data center operating in FC



(a) FC mode

(b) FMC mode

Figure 12: Comparison of regulation capacities at different cooling modes

mode and FMC mode with $\beta = 1.1$, $\tau = 5$ min and RegA signal. In FC mode, the data center can provide almost same regulation capacity using s_1 and s_2. To achieve regulation down capacity, servers work at the allowable minimum frequency in both strategies, thus consume the same amount of power for a given workload. However, the CHWST is reset to 10 °C in s_2 instead of 8 °C in s_1. To address the same cooling load, the higher the CHWST is, the more power is consumed by chilled water pumps and CRAH fans to maintain the same supply air temperature and room temperature, and the less power can be consumed by cooling tower fans because of the CHWST control logic during FC mode described in Section 4.1. The constant-speed condenser water pumps consume the same amount of power in the two strategies. The tradeoff among the power changes of cooling towers, pumps, and CRAH fans determines how much more or less capacity s_2 can provide compared with s_1.

In FMC mode, s_2 can provide 18 - 76 kW more regulation capacity than s_1. The minimum regulation capacity for both strategies occurs at around 4am-7am when the workload is the lightest. For example, the capacity is about 34 kW (1.2% of the nominal data center power) for s_2 and 16 kW (0.6% of the nominal data center power) for s_1 at 4am. The maximum regulation capacity is at around 9pm-11pm when the workload is the heaviest. The capacity at 10pm by s_2 is about 460 kW (17% of the data center nominal power), and that by s_1 is about 384 kW (14% of the data center nominal power). In FMC mode, s_2 outperforms s_1 in terms of regulation capacity because increasing CHWST can reduce the power consumption of the chillers. The tradeoff power changes of chillers, CRAH fans, pumps and cooling towers determines that increasing CHWST in the proposed range can reduce the cooling system power consumption, thus increase the regulation down capacity. Note that the CHWST range should be case-specific in a different cooling system based on the curve between the system efficiency and CHWST.

We also roughly estimated the cost savings from providing the FR service using the proposed synergistic strategy in 2019. The data center is considered an industrial load, and the average energy price for industrial loads in Illinois is reported as around \$ 0.066/kWh in 2019 [59]. The hourly regulation market clearing price in 2019 can be downloaded from PJM database [40]. Here we assume the same daily workload profile for the whole year, and hourly capacity bid as calculated by the model perturbation method is always accepted in the regulation market. The estimated annual energy costs for the data center is \$745,910, and the annual revenues from providing regulation service in PJM is \$29,589, which saves 4.0% of energy costs compared with the same data center without providing FR service.

In summary, the proposed synergistic control strategy can better extract the extra potentials of the cooling system together with servers to provide FR service for power grids, especially when chillers are activated. Without chillers, the cooling system consumes a small amount of power, and resetting CHWST can only provide an insignificant regulation capacity.

6. Conclusions

In this paper, we proposed a synergistic control strategy for data centers to provide frequency regulation service in an electric market. We also developed a flexibility factor which has been shown to improve regulation performance scores. To fully investigate the important factors that can affect the frequency regulation performance, a case study that sweeps different parameters with different values at a whole-system level was conducted. Simulation results showed that the data center can provide a regulation capacity as large as 17% of its nominal power.

Furthermore, the performance of the data center providing frequency regulation service for both RegA and RegD using the proposed strategy is largely influenced by the bidding regulation capacity C_{reg}, frequency regulation flexibility factor β, workload condition and cooling mode of the cooling system, and minimally influenced by the time constant of chillers τ. The main findings of the paper can be summarized as follows:

- The larger the bidding regulation capacity is, the worse the frequency regulation service performance is. The performance is degraded mainly because of the insufficient regulation down capacity, which is determined by the minimum aggregated frequency and the number of active servers.

- The proposed flexibility factor β can improve the frequency regulation service when tuned appropriately according to the coefficients of the aggregator model. In this paper, the preferable lower limit is 1.0, and the upper limit should be decided by considering the power tradeoff between more activated servers and increased minimum aggregated frequency.
- Thermal time constant of the chillers τ has minimal influence on the provided frequency regulation service, because the fast-response resources (e.g., servers) can compensate the delays caused by the slow-response resources (e.g., chillers).
- The larger the workload is, with the same amount of bidding capacity, the better the regulation performance is.
- The system can provide better frequency regulation service in Fully Mechanical Cooling mode than Free Cooling mode, because larger regulation down capacity can be provided in the former.

In the end, this paper compares the frequency regulation capacity the data center can provide with different control strategies. The proposed control strategy can provide an extra regulation (76 kW) of 3% of the design power in the data center during Fully Mechanical Cooling mode compared with a server-only control strategy, while in Free Cooling mode these two strategies have almost the same regulation capacity. The capacity difference is mainly caused by the tradeoff of power changes among different cooling equipment in different operational conditions. The estimated annual revenues from the regulation market using the proposed synergistic strategy is $29,589, 4% of relative saving compared with the same data center without providing frequency regulation service.

7. Acknowledgment

This material is based upon work supported by the U.S. Department of Energys Office of Energy Efficiency and Renewable Energy (EERE) under the Building Technologies Office Award Number DE-EE0007688.

8. Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

9. Appendix

9.1. Mathematical Models of Data Center Systems

Here, we provide additional details regarding the models for the systems within a data center.

9.1.1. IT equipment

An aggregated server model described in Ref [34] is adopted here. This model can output the real-time power and service response time based on CPU frequency, workload, and number of active servers.

$$P_{servers}(t) = \lambda(t) \sum_{0}^{r} b_i f_{agg}(t)^i + \sum_{0}^{s} c_j N_{act}(t)^j, 0 \leq i \leq r, 0 \leq j \leq s \tag{22}$$

where b_i, c_j are constant coefficients that can be obtained from curve fitting techniques.

Here we use the average response time to quantify the service quality of a data center. The workloads are modeled as GI/G/m queues, which assumes a general distribution with independent arrival times and a general distribution of service times. The total time that a job spends in the queuing system is known as response time. The response time usually consists of two parts: waiting time, that is, the time that a job spends in a queue waiting to be serviced; and service time, that is, the time that a job needs to be executed. The average response time model is adopted from [60]. Details are shown as follows.

$$\mu(t) = k f_{agg}(t) \tag{23}$$

$$t_s = \frac{1}{\mu(t)} \tag{24}$$

$$\rho(t) = \frac{\lambda(t)}{N_{act}(t)\mu(t)}, 0 \leq \rho(t) \leq 1 \tag{25}$$

$$P_m = \begin{cases} \frac{\rho(t)^{N_{act}(t)}+\rho(t)}{2}, & \rho(t) \geq 0.7 \\ \rho(t)^{\frac{N_{act}(t)+1}{2}}, & \rho(t) < 0.7 \end{cases} \tag{26}$$

$$t_w = \frac{C_A^2 + C_B^2}{2N_{act}(t)} \frac{P_m}{\mu(t)(1-\rho(t))} \tag{27}$$

$$t_r = t_s + t_w \tag{28}$$

In the above equations, ρ is the average utilization of the server, representing the fraction of occupied time, P_m is approximated probability that an arriving job is queued, t_w is the waiting time, C_A and C_B are constant coefficients reflecting the type of data centers.

9.1.2. Dynamics in Heat Transfer

The cooling system is modeled using Modelica Buildings library to leverage the latest development of data center models in [52, 53, 16]. The Modelica-based models are able to capture thermal and mechanical dynamics in the system. Details are explained as follows.

Instead of using steady state energy conservation equations, we use the following equation to quantify the dynamic thermal response in the heat transfer units:

$$mC_p\frac{\mathrm{d}T}{\mathrm{d}t} = \dot{m}_iC_pT_i - \dot{m}_oC_pT + \dot{q} \tag{29}$$

By assuming steady state mass conservation, we get

$$\dot{m}_i = \dot{m}_o = \dot{m} \tag{30}$$

Replacing Eq. (30) into Eq. (29), we get

$$\frac{m}{\dot{m}}\frac{\mathrm{d}T}{\mathrm{d}t} + T = T_i + \frac{\dot{q}}{\dot{m}C_p} \tag{31}$$

Therefore, the time constant of the heat transfer unit is

$$\tau = \frac{m}{\dot{m}} \tag{32}$$

In the above equations, m is the mass of the fluid volume, $\dot{m}_i$ and $\dot{m}_o$ are the mass flow rates at inlet and outlet, T is the temperature of the fluid volume, T_i and T_o are the temperature at the volume inlet and outlet, $\dot{q}$ is the heat rate transferred into the fluid volume, and C_p is the fluid specific heat capacity.

9.1.3. Dynamics in Equipment On/Off

The dynamics during equipment start or shutdown are represented by a second-order critical damping low pass filter. Detailed equations are listed as below. a is the coefficient of the state space equations for the first-order filter described in Eq. (34) and Eq. (35). f_{cut} is the cut-off frequency of the low pass filter, which passes signals with a frequency lower than f_{cut} and attenuates signals with a higher frequency. α is a frequency correction factor for different orders. Here we use 0.622 for the second order. u is the input signal passed to the filter. $x_{1,2}$ are the filtered signals, and y is the output signal from the filter.

$$a = -\frac{2\pi f_{cut}}{\alpha} \tag{33}$$

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = ax_1 - au \tag{34}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = ax_2 - ax_1 \tag{35}$$

$$y = x_2 \tag{36}$$

9.2. Performance Scores for Different Scenarios

Table 3: Performance for RegA test signal during light load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.6536	0.5749	0.3861	0.5382
		10	0.6509	0.5857	0.3853	0.5406
		15	0.6362	0.6125	0.3864	0.5450
	1.0	5	0.8506	0.7382	0.6439	0.7442
		10	0.8475	0.7292	0.6435	0.7401
		15	0.8479	0.7224	0.6456	0.7386
	1.1	5	0.9149	0.8150	0.7614	0.8304
		10	0.9148	0.8049	0.7592	0.8263
		15	0.9129	0.8210	0.7598	0.8312
268.0	0.9	5	0.6542	0.5815	0.3322	0.5226
		10	0.6425	0.6286	0.3318	0.5343
		15	0.6500	0.5618	0.3315	0.5144
	1.0	5	0.7231	0.6886	0.4702	0.6273
		10	0.7008	0.5936	0.4703	0.5882
		15	0.7037	0.5885	0.4723	0.5882
	1.1	5	0.7967	0.7068	0.5455	0.6830
		10	0.8020	0.7382	0.5433	0.6945
		15	0.8040	0.7410	0.5415	0.6955
402.0	0.9	5	0.6353	0.5606	0.3187	0.5048
		10	0.6445	0.5724	0.3173	0.5114
		15	0.6498	0.5331	0.3187	0.5005
	1.0	5	0.6717	0.6169	0.4120	0.5669
		10	0.6770	0.6035	0.4109	0.5638
		15	0.6864	0.5817	0.4112	0.5597
	1.1	5	0.7124	0.6144	0.4651	0.5973
		10	0.6964	0.6435	0.4647	0.6015
		15	0.7136	0.6354	0.4632	0.6041

Table 4: Performance for RegA test signal during medium load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8258	0.4914	0.4627	0.5933
		10	0.8280	0.4957	0.4604	0.5947
		15	0.8280	0.4954	0.4603	0.5946
	1.0	5	0.9987	1.0000	0.9469	0.9819
		10	0.9987	1.0000	0.9475	0.9821
		15	0.9987	1.0000	0.9472	0.9820
	1.1	5	1.0000	1.0000	0.9475	0.9825
		10	1.0000	1.0000	0.9474	0.9825
		15	1.0000	1.0000	0.9473	0.9824
268.0	0.9	5	0.8168	0.4972	0.4084	0.5741
		10	0.8170	0.4938	0.4082	0.5730
		15	0.8145	0.4897	0.4088	0.5710
	1.0	5	0.9063	0.8229	0.7694	0.8329
		10	0.9061	0.8210	0.7699	0.8323
		15	0.9063	0.8199	0.7701	0.8321
	1.1	5	0.9774	0.9371	0.8784	0.9310
		10	0.9786	0.9342	0.8787	0.9305
		15	0.9776	0.9375	0.8788	0.9313
402.0	0.9	5	0.8121	0.4988	0.3819	0.5642
		10	0.8120	0.4988	0.3819	0.5642
		15	0.8119	0.5039	0.3817	0.5658
	1.0	5	0.8392	0.7132	0.6391	0.7305
		10	0.8404	0.7126	0.6390	0.7307
		15	0.8392	0.7135	0.6391	0.7306
	1.1	5	0.9600	0.8056	0.7428	0.8361
		10	0.9602	0.8051	0.7406	0.8353
		15	0.9600	0.8051	0.7407	0.8353

Table 5: Performance for RegA test signal during heavy load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8103	0.4501	0.3822	0.5476
		10	0.8100	0.4499	0.3822	0.5474
		15	0.8098	0.4499	0.3819	0.5472
	1.0	5	1.0000	1.0000	0.9474	0.9825
		10	1.0000	1.0000	0.9492	0.9831
		15	1.0000	1.0000	0.9494	0.9831
	1.1	5	1.0000	1.0000	0.9471	0.9824
		10	1.0000	1.0000	0.9473	0.9824
		15	1.0000	1.0000	0.9475	0.9825
268.0	0.9	5	0.8286	0.4501	0.3706	0.5498
		10	0.8283	0.4503	0.3707	0.5497
		15	0.8286	0.4499	0.3707	0.5497
	1.0	5	0.9847	0.9508	0.9126	0.9494
		10	0.9846	0.9508	0.9126	0.9493
		15	0.9846	0.9511	0.9107	0.9488
	1.1	5	1.0000	1.0000	0.9471	0.9824
		10	1.0000	1.0000	0.9470	0.9823
		15	1.0000	1.0000	0.9490	0.9830
402.0	0.9	5	0.8366	0.4526	0.3665	0.5519
		10	0.8366	0.4529	0.3663	0.5520
		15	0.8364	0.4526	0.3668	0.5520
	1.0	5	0.9714	0.8850	0.8017	0.8860
		10	0.9714	0.8875	0.7995	0.8861
		15	0.9714	0.8875	0.7998	0.8862
	1.1	5	0.9983	0.9989	0.9454	0.9809
		10	0.9983	0.9989	0.9468	0.9813
		15	0.9983	0.9989	0.9455	0.9809

Table 6: Performance for RegA test signal during light load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9869	0.9431	0.4293	0.7864
		10	0.9879	0.9351	0.4346	0.7859
		15	0.9869	0.9404	0.4310	0.7861
	1.0	5	0.9703	0.8682	0.7080	0.8488
		10	0.9726	0.8761	0.7062	0.8516
		15	0.9712	0.8467	0.7053	0.8411
	1.1	5	0.9879	0.9051	0.8160	0.9030
		10	0.9877	0.9011	0.8160	0.9016
		15	0.9870	0.9021	0.8135	0.9009
268.0	0.9	5	0.9788	0.9468	0.3557	0.7604
		10	0.9802	0.9424	0.3556	0.7594
		15	0.9806	0.9432	0.3550	0.7596
	1.0	5	0.9735	0.9187	0.5078	0.8000
		10	0.9722	0.8940	0.5084	0.7915
		15	0.9672	0.9146	0.5074	0.7964
	1.1	5	0.9409	0.9104	0.5747	0.8087
		10	0.9418	0.9064	0.5728	0.8070
		15	0.9431	0.8875	0.5740	0.8015
402.0	0.9	5	0.9751	0.9463	0.3318	0.7511
		10	0.9759	0.9426	0.3310	0.7498
		15	0.9767	0.9424	0.3307	0.7499
	1.0	5	0.9758	0.9144	0.4329	0.7744
		10	0.9754	0.9149	0.4325	0.7743
		15	0.9763	0.9164	0.4330	0.7752
	1.1	5	0.9694	0.9211	0.4809	0.7905
		10	0.9694	0.9188	0.4810	0.7897
		15	0.9690	0.9165	0.4813	0.7890

Table 7: Performance for RegA test signal during medium load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9035	0.9015	0.4903	0.7651
		10	0.9254	0.9336	0.4839	0.7810
		15	0.9278	0.9296	0.5109	0.7894
	1.0	5	1.0000	1.0000	0.9446	0.9815
		10	1.0000	1.0000	0.9477	0.9826
		15	1.0000	1.0000	0.9481	0.9827
	1.1	5	1.0000	1.0000	0.9452	0.9817
		10	1.0000	1.0000	0.9493	0.9831
		15	1.0000	1.0000	0.9459	0.9820
268.0	0.9	5	0.9030	0.9086	0.4157	0.7424
		10	0.9330	0.9421	0.4208	0.7653
		15	0.9339	0.9556	0.4242	0.7712
	1.0	5	0.9370	0.9243	0.8400	0.9004
		10	0.9461	0.9282	0.8449	0.9064
		15	0.9516	0.9339	0.8431	0.9096
	1.1	5	0.9767	0.9861	0.9362	0.9663
		10	0.9778	0.9856	0.9341	0.9658
		15	0.9788	0.9857	0.9330	0.9658
402.0	0.9	5	0.9024	0.9081	0.3954	0.7353
		10	0.9285	0.9493	0.3994	0.7591
		15	0.9215	0.9256	0.4011	0.7494
	1.0	5	0.9045	0.8418	0.6906	0.8123
		10	0.9164	0.8286	0.6961	0.8137
		15	0.9206	0.8311	0.6940	0.8152
	1.1	5	0.9475	0.9043	0.8103	0.8874
		10	0.9569	0.9114	0.8139	0.8941
		15	0.9643	0.9172	0.8129	0.8981

Table 8: Performance for RegA test signal during heavy load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8226	0.9308	0.3526	0.7020
		10	0.9061	0.9661	0.3405	0.7376
		15	0.9356	0.9846	0.3597	0.7600
	1.0	5	1.0000	1.0000	0.9480	0.9827
		10	1.0000	1.0000	0.9479	0.9826
		15	1.0000	1.0000	0.9459	0.9820
	1.1	5	1.0000	1.0000	0.9473	0.9824
		10	1.0000	1.0000	0.9495	0.9832
		15	1.0000	1.0000	0.9487	0.9829
268.0	0.9	5	0.8225	0.9246	0.3456	0.6976
		10	0.9050	0.9607	0.3426	0.7361
		15	0.9314	0.9747	0.3575	0.7545
	1.0	5	1.0000	1.0000	0.9465	0.9822
		10	1.0000	1.0000	0.9502	0.9834
		15	1.0000	1.0000	0.9445	0.9815
	1.1	5	1.0000	1.0000	0.9460	0.9820
		10	1.0000	1.0000	0.9484	0.9828
		15	1.0000	1.0000	0.9482	0.9827
402.0	0.9	5	0.8395	0.9537	0.3465	0.7132
		10	0.9150	0.9637	0.3538	0.7442
		15	0.9367	0.9686	0.3602	0.7552
	1.0	5	0.9444	0.9624	0.8906	0.9324
		10	0.9621	0.9733	0.9047	0.9467
		15	0.9744	0.9792	0.9087	0.9541
	1.1	5	1.0000	1.0000	0.9483	0.9828
		10	1.0000	1.0000	0.9479	0.9826
		15	1.0000	1.0000	0.9498	0.9833

Table 9: Performance for RegD test signal during light load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9041	0.8353	0.5876	0.7757
		10	0.9006	0.8325	0.5867	0.7733
		15	0.9045	0.8354	0.5882	0.7760
	1.0	5	0.9432	0.8467	0.6744	0.8214
		10	0.9445	0.8471	0.6740	0.8219
		15	0.9445	0.8468	0.6748	0.8220
	1.1	5	0.9466	0.8669	0.7405	0.8514
		10	0.9465	0.8628	0.7399	0.8497
		15	0.9460	0.8708	0.7399	0.8523
268.0	0.9	5	0.7792	0.8156	0.4327	0.6758
		10	0.7777	0.8129	0.4328	0.6745
		15	0.7800	0.8047	0.4331	0.6726
	1.0	5	0.8004	0.8172	0.4823	0.7000
		10	0.8011	0.8168	0.4819	0.6999
		15	0.7981	0.8125	0.4811	0.6973
	1.1	5	0.8021	0.8263	0.5043	0.7109
		10	0.8011	0.8210	0.5049	0.7090
		15	0.8032	0.8149	0.5032	0.7071
402.0	0.9	5	0.7524	0.7978	0.3622	0.6375
		10	0.7506	0.7792	0.3636	0.6311
		15	0.7532	0.7931	0.3633	0.6365
	1.0	5	0.7657	0.7957	0.3979	0.6531
		10	0.7663	0.7894	0.3974	0.6511
		15	0.7696	0.7933	0.3980	0.6537
	1.1	5	0.7709	0.7946	0.4135	0.6597
		10	0.7736	0.7958	0.4131	0.6608
		15	0.7737	0.7978	0.4135	0.6617

Table 10: Performance for RegD test signal during medium load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9175	0.8179	0.6286	0.7880
		10	0.9175	0.8179	0.6286	0.7880
		15	0.9176	0.8179	0.6286	0.7880
	1.0	5	1.0000	1.0000	0.8634	0.9545
		10	1.0000	1.0000	0.8636	0.9545
		15	1.0000	1.0000	0.8636	0.9545
	1.1	5	1.0000	1.0000	0.8644	0.9548
		10	1.0000	1.0000	0.8644	0.9548
		15	1.0000	1.0000	0.8644	0.9548
268.0	0.9	5	0.9185	0.8232	0.6132	0.7850
		10	0.9185	0.8232	0.6132	0.7850
		15	0.9185	0.8232	0.6132	0.7849
	1.0	5	0.9530	0.8500	0.7476	0.8502
		10	0.9531	0.8500	0.7477	0.8503
		15	0.9530	0.8500	0.7480	0.8503
	1.1	5	0.9861	1.0000	0.8222	0.9361
		10	0.9861	1.0000	0.8218	0.9360
		15	0.9860	1.0000	0.8221	0.9360
402.0	0.9	5	0.9174	0.8272	0.5819	0.7755
		10	0.9176	0.8272	0.5819	0.7756
		15	0.9175	0.8272	0.5820	0.7756
	1.0	5	0.9471	0.8417	0.6680	0.8189
		10	0.9473	0.8417	0.6680	0.8190
		15	0.9472	0.8417	0.6680	0.8190
	1.1	5	0.9488	0.8500	0.7232	0.8407
		10	0.9488	0.8500	0.7232	0.8407
		15	0.9489	0.8500	0.7232	0.8407

Table 11: Performance for RegD test signal during heavy load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8850	0.7857	0.6005	0.7571
		10	0.8850	0.7857	0.6005	0.7571
		15	0.8852	0.7857	0.6006	0.7572
	1.0	5	1.0000	1.0000	0.8635	0.9545
		10	1.0000	1.0000	0.8636	0.9545
		15	1.0000	1.0000	0.8635	0.9545
	1.1	5	1.0000	1.0000	0.8635	0.9545
		10	1.0000	1.0000	0.8635	0.9545
		15	1.0000	1.0000	0.8636	0.9545
268.0	0.9	5	0.9050	0.7911	0.5984	0.7648
		10	0.9050	0.7911	0.5984	0.7648
		15	0.9050	0.7911	0.5983	0.7648
	1.0	5	0.9885	1.0000	0.8334	0.9406
		10	0.9884	1.0000	0.8333	0.9406
		15	0.9885	1.0000	0.8334	0.9406
	1.1	5	1.0000	1.0000	0.8639	0.9546
		10	1.0000	1.0000	0.8639	0.9546
		15	1.0000	1.0000	0.8639	0.9546
402.0	0.9	5	0.9129	0.7988	0.5957	0.7691
		10	0.9131	0.7989	0.5956	0.7692
		15	0.9131	0.7989	0.5956	0.7692
	1.0	5	0.9536	0.8625	0.7621	0.8594
		10	0.9536	0.8625	0.7617	0.8593
		15	0.9537	0.8625	0.7619	0.8594
	1.1	5	0.9988	1.0000	0.8585	0.9524
		10	0.9988	1.0000	0.8583	0.9524
		15	0.9988	1.0000	0.8584	0.9524

Table 12: Performance for RegD test signal during light load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9530	0.9940	0.6341	0.8604
		10	0.9550	0.9935	0.6305	0.8597
		15	0.9547	0.9943	0.6338	0.8610
	1.0	5	0.9615	0.9064	0.7398	0.8692
		10	0.9612	0.9444	0.7379	0.8812
		15	0.9620	0.9522	0.7370	0.8837
	1.1	5	0.9667	0.9947	0.7949	0.9188
		10	0.9667	0.9989	0.7922	0.9193
		15	0.9654	0.9989	0.7924	0.9189
268.0	0.9	5	0.8972	0.9628	0.4784	0.7795
		10	0.8991	0.9824	0.4790	0.7868
		15	0.8906	0.9760	0.4779	0.7815
	1.0	5	0.9003	0.9304	0.5275	0.7861
		10	0.9021	0.9390	0.5274	0.7895
		15	0.9017	0.9425	0.5271	0.7904
	1.1	5	0.9015	0.9276	0.5478	0.7923
		10	0.9016	0.9562	0.5461	0.8013
		15	0.9013	0.9667	0.5468	0.8049
402.0	0.9	5	0.9080	0.9599	0.3887	0.7522
		10	0.9122	0.9608	0.3886	0.7539
		15	0.9133	0.9843	0.3899	0.7625
	1.0	5	0.9042	0.9196	0.4242	0.7493
		10	0.9083	0.9254	0.4226	0.7521
		15	0.9084	0.9326	0.4237	0.7549
	1.1	5	0.9001	0.9050	0.4371	0.7474
		10	0.9050	0.9247	0.4369	0.7556
		15	0.8520	0.9325	0.4377	0.7408

Table 13: Performance for RegD test signal during medium load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9615	0.9992	0.6875	0.8827
		10	0.9726	0.9994	0.6906	0.8875
		15	0.9769	0.9988	0.6891	0.8883
	1.0	5	1.0000	1.0000	0.8642	0.9547
		10	1.0000	1.0000	0.8648	0.9549
		15	1.0000	1.0000	0.8649	0.9550
	1.1	5	1.0000	1.0000	0.8648	0.9549
		10	1.0000	1.0000	0.8648	0.9549
		15	1.0000	1.0000	0.8652	0.9551
268.0	0.9	5	0.9491	0.9986	0.6364	0.8614
		10	0.9599	0.9983	0.6375	0.8653
		15	0.9649	0.9983	0.6390	0.8674
	1.0	5	0.9792	0.9999	0.8067	0.9286
		10	0.9801	0.9997	0.8059	0.9286
		15	0.9790	0.9996	0.8071	0.9286
	1.1	5	0.9991	1.0000	0.8589	0.9527
		10	0.9991	1.0000	0.8559	0.9517
		15	0.9992	1.0000	0.8595	0.9529
402.0	0.9	5	0.9436	0.9978	0.6226	0.8547
		10	0.9568	0.9979	0.6252	0.8600
		15	0.9596	0.9974	0.6247	0.8605
	1.0	5	0.9597	0.9657	0.7267	0.8840
		10	0.9617	0.9739	0.7278	0.8878
		15	0.9643	0.9738	0.7262	0.8881
	1.1	5	0.9716	0.9997	0.7952	0.9222
		10	0.9731	0.9992	0.7934	0.9219
		15	0.9722	0.9992	0.7940	0.9218

Table 14: Performance for RegD test signal during heavy load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9579	0.9999	0.7003	0.8860
		10	0.9808	1.0000	0.7674	0.9161
		15	0.9894	1.0000	0.7814	0.9236
	1.0	5	1.0000	1.0000	0.8657	0.9552
		10	1.0000	1.0000	0.8644	0.9548
		15	1.0000	1.0000	0.8657	0.9552
	1.1	5	1.0000	1.0000	0.8663	0.9554
		10	1.0000	1.0000	0.8671	0.9557
		15	1.0000	1.0000	0.8659	0.9553
268.0	0.9	5	0.9479	0.9996	0.6357	0.8611
		10	0.9717	0.9990	0.6548	0.8752
		15	0.9817	0.9990	0.6587	0.8798
	1.0	5	1.0000	1.0000	0.8645	0.9548
		10	1.0000	1.0000	0.8650	0.9550
		15	1.0000	1.0000	0.8642	0.9547
	1.1	5	1.0000	1.0000	0.8656	0.9552
		10	1.0000	1.0000	0.8650	0.9550
		15	1.0000	1.0000	0.8644	0.9548
402.0	0.9	5	0.9439	0.9988	0.6164	0.8530
		10	0.9665	0.9967	0.6272	0.8634
		15	0.9762	0.9985	0.6332	0.8693
	1.0	5	0.9955	1.0000	0.8430	0.9461
		10	0.9967	1.0000	0.8503	0.9490
		15	0.9971	1.0000	0.8498	0.9490
	1.1	5	1.0000	1.0000	0.8636	0.9545
		10	1.0000	1.0000	0.8645	0.9548
		15	1.0000	1.0000	0.8650	0.9550

10. References

[1] J. Koomey, Growth in data center electricity use 2005 to 2010, A report by Analytical Press, completed at the request of The New York Times 9.

[2] R. Brown, et al., Report to congress on server and data center energy efficiency: Public law 109-431.

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

[4] ASHRAE, Thermal guidelines for data processing environments, 4th Edition, 2015.

[5] L. Chen, N. Li, On the interaction between load balancing and speed scaling, IEEE Journal on Selected Areas in Communications 33 (12) (2015) 2567–2578.

[6] Q. Zhang, M. F. Zhani, S. Zhang, Q. Zhu, R. Boutaba, J. L. Hellerstein, Dynamic energy-aware capacity provisioning for cloud computing environments, in: Proceedings of the 9th international conference on Autonomic computing, ACM, 2012, pp. 145–154.

[7] H. Chen, C. Hankendi, M. C. Caramanis, A. K. Coskun, Dynamic server power capping for enabling data center participation in power markets, in: 2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), IEEE, 2013, pp. 122–129.

[8] M. Lin, A. Wierman, L. L. Andrew, E. Thereska, Dynamic right-sizing for power-proportional data centers, IEEE/ACM Transactions on Networking (TON) 21 (5) (2013) 1378–1391.

[9] J. Heo, P. Jayachandran, I. Shin, D. Wang, T. Abdelzaher, X. Liu, Optituner: On performance composition and server farm energy minimization

application, IEEE Transactions on Parallel and Distributed Systems 22 (11) (2011) 1871–1878.

[10] L. Zhang, S. Ren, C. Wu, Z. Li, A truthful incentive mechanism for emergency demand response in colocation data centers, in: 2015 IEEE Conference on Computer Communications (INFOCOM), IEEE, 2015, pp. 2632–2640.

[11] V. Ganti, G. Ghatikar, Smart grid as a driver for energy-intensive industries: a data center case study, Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2012).

[12] Y. Yao, L. Huang, A. Sharma, L. Golubchik, M. Neely, Data centers power reduction: A two time scale approach for delay tolerant workloads, in: 2012 Proceedings IEEE INFOCOM, IEEE, 2012, pp. 1431–1439.

[13] T. I. A. Standards, T. Dept, A. N. S. Institute, Telecommunications Infrastructure Standard for Data Centers, Telecommunication Industry Association, 2005.

[14] G. Ghatikar, M. A. Piette, S. Fujita, A. McKane, J. H. Dudley, A. Radspieler, K. Mares, D. Shroyer, Demand response and open automated demand response opportunities for data centers, Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2009).

[15] N. Bates, G. Ghatikar, G. Abdulla, G. A. Koenig, S. Bhalachandra, M. Sheikhalishahi, T. Patki, B. Rountree, S. Poole, Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges, Informatik-Spektrum 38 (2) (2015) 111–127.

[16] Y. Fu, W. Zuo, M. Wetter, J. W. VanGilder, P. Yang, Equation-based object-oriented modeling and simulation of data center cooling systems, Energy and Buildings 198 (2019) 503 – 519. doi:https://doi.org/10.1016/j.enbuild.2019.06.037.

URL http://www.sciencedirect.com/science/article/pii/S0378778819307078

[17] Y.-J. Kim, L. K. Norford, J. L. Kirtley, Modeling and analysis of a variable speed heat pump for frequency regulation through direct load control, IEEE Transactions on Power Systems 30 (1) (2015) 397–408.

[18] X. Xue, S. Wang, C. Yan, B. Cui, A fast chiller power demand response control strategy for buildings connected to smart grid, Applied Energy 137 (2015) 77–87.

[19] L. Su, L. K. Norford, Demonstration of hvac chiller control for power grid frequency regulation-part 1: Controller development and experimental results, Science and Technology for the Built Environment 21 (8) (2015) 1134–1142.

[20] L. Su, L. K. Norford, Demonstration of hvac chiller control for power grid frequency regulation-part 2: Discussion of results and considerations for broader deployment, Science and Technology for the Built Environment 21 (8) (2015) 1143–1153.

[21] G. Ghatikar, V. Ganti, N. Matson, M. A. Piette, Demand response opportunities and enabling technologies for data centers: Findings from field studies, Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2012).

[22] J. MacDonald, S. Kiliccote, J. Boch, J. Chen, R. Nawy, Commercial building loads providing ancillary services in pjm, Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2014).

[23] H. Hao, A. Kowli, Y. Lin, P. Barooah, S. Meyn, Ancillary service for the grid via control of commercial building hvac systems, in: 2013 American control conference, IEEE, 2013, pp. 467–472.

[24] H. Hao, Y. Lin, A. S. Kowli, P. Barooah, S. Meyn, Ancillary service to the grid through control of fans in commercial building hvac systems, IEEE Transactions on smart grid 5 (4) (2014) 2066–2074.

[25] Y. Lin, P. Barooah, S. Meyn, T. Middelkoop, Experimental evaluation of frequency regulation from commercial building hvac systems, IEEE Transactions on Smart Grid 6 (2) (2015) 776–783.

[26] E. Vrettos, E. C. Kara, J. MacDonald, G. Andersson, D. S. Callaway, Experimental demonstration of frequency regulation by commercial buildings-part i: Modeling and hierarchical control design, IEEE Transactions on Smart Grid 9 (4) (2018) 3213–3223.

[27] E. Vrettos, E. C. Kara, J. MacDonald, G. Andersson, D. S. Callaway, Experimental demonstration of frequency regulation by commercial buildings-part ii: results and performance evaluation, IEEE Transactions on Smart Grid 9 (4) (2018) 3224–3234.

[28] I. Beil, I. Hiskens, S. Backhaus, Frequency regulation from commercial building hvac demand response, Proceedings of the IEEE 104 (4) (2016) 745–757.

[29] M. Maasoumy, J. Ortiz, D. Culler, A. Sangiovanni-Vincentelli, Flexibility of commercial building hvac fan as ancillary service for smart grid, arXiv preprint arXiv:1311.6094.

[30] Y. Lin, P. Barooah, S. Meyn, T. Middelkoop, Demand side frequency regulation from commercial building hvac systems: An experimental study, in: 2015 American Control Conference (ACC), IEEE, 2015, pp. 3019–3024.

[31] P. Zhao, G. P. Henze, S. Plamp, V. J. Cushing, Evaluation of commercial building hvac systems as frequency regulation providers, Energy and Buildings 67 (2013) 225–235.

[32] Z. Liu, A. Wierman, Y. Chen, B. Razon, N. Chen, Data center demand response: Avoiding the coincident peak via workload shifting and local generation, Performance Evaluation 70 (10) (2013) 770–791.

[33] J. Li, Z. Li, K. Ren, X. Liu, Towards optimal electric demand management for internet data centers, IEEE Transactions on Smart Grid 3 (1) (2012) 183–192.

[34] H. Chen, A. K. Coskun, M. C. Caramanis, Real-time power control of data centers for providing regulation service, in: 52nd IEEE Conference on Decision and Control, IEEE, 2013, pp. 4314–4321.

[35] W. Wang, A. Abdolrashidi, N. Yu, D. Wong, Frequency regulation service provision in data center with computational flexibility, Applied Energy 251 (2019) 113304.

[36] J. McClurg, Fast demand response with datacenter loads: A green dimension of big data, Ph.D. thesis, University of Iowa (2017).

[37] J. Li, J. F. Martinez, Dynamic power-performance adaptation of parallel computation on chip multiprocessors, in: The Twelfth International Symposium on High-Performance Computer Architecture, 2006., 2006, pp. 77–87.

[38] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, C. Le, Rapl: memory power estimation and capping, in: 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), IEEE, 2010, pp. 189–194.

[39] P. I. LLC, Pjm manual 11: Energy & ancillary services market operations (2019).

[40] PJM, Ancillary services (2019).
URL https://www.pjm.com/markets-and-operations/ancillary-services.aspx

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

[42] M. Wolf, Chapter 5 - processors and systems, in: M. Wolf (Ed.), The Physics of Computing, Morgan Kaufmann, Boston, 2017, pp. 149 – 203. doi:https://doi.org/10.1016/B978-0-12-809381-8.00005-5.
URL http://www.sciencedirect.com/science/article/pii/B9780128093818000055

[43] I. Narayanan, D. Wang, A. Sivasubramaniam, H. K. Fathy, S. James, et al., Evaluating energy storage for a multitude of uses in the datacenter, in: 2017 IEEE International Symposium on Workload Characterization (IISWC), IEEE, 2017, pp. 12–21.

[44] Y. Shi, B. Xu, B. Zhang, D. Wang, Leveraging energy storage to optimize data center electricity cost in emerging power markets, in: Proceedings of the Seventh International Conference on Future Energy Systems, ACM, 2016, p. 18.

[45] B. Aksanli, T. Rosing, Providing regulation services and managing data center peak power budgets, in: 2014 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE, 2014, pp. 1–4.

[46] H. Chen, Z. Liu, A. K. Coskun, A. Wierman, Optimizing energy storage participation in emerging power markets, in: 2015 Sixth International Green and Sustainable Computing Conference (IGSC), IEEE, 2015, pp. 1–6.

[47] Eaton launches industry first ups-as-a-reserve service to support the power grid in frequency containment reserve, http://powerquality.eaton.com/emea/about-us/news-events/2017/pr031017.asp?act=smtc&id=&key=&Quest_user_id=&leadg_Q_QRequired=&site=&menu=&cx=98&x=16&y=8, accessed: 2019-09-30.

[48] R. Guruprasad, P. Murali, D. Krishnaswamy, S. Kalyanaraman, Coupling a small battery with a datacenter for frequency regulation, in: 2017 IEEE Power & Energy Society General Meeting, IEEE, 2017, pp. 1–5.

[49] S. Li, M. Brocanelli, W. Zhang, X. Wang, Integrated power management of data centers and electric vehicles for energy and regulation market participation, IEEE Transactions on Smart Grid 5 (5) (2014) 2283–2294.

[50] M. Brocanelli, S. Li, X. Wang, W. Zhang, Joint management of data centers and electric vehicles for maximized regulation profits, in: 2013 International Green Computing Conference Proceedings, IEEE, 2013, pp. 1–10.

[51] U. Institute, Annual data center survey results, Report, Uptime Institute (2019).

[52] Y. Fu, M. Wetter, W. Zuo, Modelica models for data center cooling systems, in: 2018 Building Performance Analysis Conference and SimBuild, Chicago, Illinois, United States of America, 2018.

[53] Y. Fu, W. Zuo, M. Wetter, J. W. VanGilder, X. Han, D. Plamondon, Equation-based object-oriented modeling and simulation for data center cooling: A case study, Energy and Buildings 186 (2019) 108 – 125. doi:https://doi.org/10.1016/j.enbuild.2019.01.018.
URL http://www.sciencedirect.com/science/article/pii/S0378778818330573

[54] Y. Fu, X. Lu, W. Zuo, Modelica models for the control evaluations of chilled water system with waterside economizer, in: Proceedings of the 13th International Modelica Conference, Regensburg, Germany, March 46, 2019, no. 157, Linkping University Electronic Press, Linkpings universitet, 2019, p. 8.

[55] B. Biegel, P. Andersen, T. S. Pedersen, K. M. Nielsen, J. Stoustrup, L. H. Hansen, Smart grid dispatch strategy for on/off demand-side devices, in: 2013 European Control Conference (ECC), IEEE, 2013, pp. 2541–2548.

[56] B. Biegel, P. Andersen, J. Stoustrup, M. B. Madsen, L. H. Hansen, L. H. Rasmussen, et al., Aggregation and control of flexible consumers–a real life demonstration, IFAC Proceedings Volumes 47 (3) (2014) 9950–9955.

[57] E. Vrettos, G. Andersson, Scheduling and provision of secondary frequency reserves by aggregations of commercial buildings, IEEE Transactions on Sustainable Energy 7 (2) (2015) 850–864.

[58] G. S. Pavlak, G. P. Henze, V. J. Cushing, Optimizing commercial building participation in energy and ancillary service markets, Energy and Buildings 81 (2014) 115–126.

[59] Table 5.6.a. average price of electricity to ultimate customers by end-use sector, https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a, accessed: 2020-07-17.

[60] G. Bolch, S. Greiner, H. De Meer, K. S. Trivedi, Queueing networks and Markov chains: modeling and performance evaluation with computer science applications, John Wiley & Sons, 2006.