Opportunities and Challenges for Data Center Demand Response

Adam Wierman Zhenhua Liu Iris Liu
Computing and Mathematical Sciences
California Institute of Technology

Hamed Mohsenian-Rad
Electrical Engineering
University of California at Riverside

***Abstract*—This paper surveys the opportunities and challenges in an emerging area of research that has the potential to significantly ease the incorporation of renewable energy into the grid as well as electric power peak-load shaving: data center demand response. Data center demand response sits at the intersection of two growing fields: energy efficient data centers and demand response in the smart grid. As such, the literature related to data center demand response is sprinkled across multiple areas and worked on by diverse groups. Our goal in this survey is to demonstrate the potential of the field while also summarizing the progress that has been made and the challenges that remain.**

I. Introduction

This paper surveys the opportunities and challenges in the emerging area of *data center demand response*. Data center demand response sits at the intersection of two important societal challenges. First, as ICT becomes increasingly crucial to society, the associated energy demands are skyrocketing, e.g., within the US the growth in electricity demand of ICT is ten times larger than the overall growth of electricity demands [1], [40], [53]. Second, the integration of renewable energy into the power grid is fundamental for improving sustainability, but causes significant challenges for management of the grid that have the potential to increase costs considerably [30], [74]. Further, this challenge is magnified by the fact that large-scale fast-charging storage is simply not cost-effective at this point.

The key idea behind data center demand response is that these two challenges are in fact symbiotic. Specifically, data centers are large loads, but are also flexible – data center loads can often be shifted in time [19], [34], [46], [58], [59], [67], [107], [109], curtailed via quality degradation [8], [45], [96], [104], or even shifted geographically [10], [57], [60], [61], [79], [81], [99], [103]. If the flexibility of data centers can be called on by the grid via demand response programs, then they can be a crucial tool for easing the incorporation of renewable energy into the grid. Further, there is potential for this interaction to be made "win-win" because the financial benefits from data center participation in demand response programs can help ease the burden from the costs of skyrocketing energy usage.

Unfortunately, despite wide recognition of the demand response potential of data centers, the current reality is that data centers perform little, if any, demand response [40], [73]. There are many reasons for this, but perhaps the biggest is simply that the demand response programs that exist today are not suited for the load profile and risk tolerance of data centers, for which availability and performance are crucial concerns.

Consequently, there is much work to be done before the true potential of data center demand response can be realized. The research ahead is highly challenging and interdisciplinary, e.g., requiring work on the management of data center participation in demand response programs and the design of new demand response markets as well as providing tools for the integration of data centers into power system modeling.

The goal of this survey is to provide a broad overview of the area of data center demand response in order to facilitate the entrance of new researchers into this interdisciplinary, important area. To that end, the survey focuses on four key issues. First, in Section II, we formally quantify the potential of data center demand response. Second, in Section III, we survey opportunities for data center participation in demand response markets that are available today. Then, in Section IV, we highlight some of the key challenges that prevent wide-scale data center participation in the existing demand response markets. Finally, in Section V, we survey some recent progress toward overcoming these challenges.

II. The potential value of data center demand response

The coming decades promise explosive growth in the use of renewable energy. For example, while the current installed capacity of wind power in the U.S. is less than 5% of total generation [27], the Department of Energy has set a goal to procure 20% of the total generation from wind power by 2030 [25]. This degree of renewable penetration brings with it major challenges for management and control of the electricity grid as a result of the unpredictable, highly variable nature of renewable energy sources.

Often, when people think of the challenges for grid management that result from increasing adoption of renewable energy, the thought is: "if only we had large-scale energy storage..." Large-scale energy storage, indeed, would solve many of the challenges associated with the unpredictability and intermittency of wind and solar energy. However, the problem is that large-scale storage is too expensive, at least for now.

It is this expense that leads to the consideration of demand response as the next-best option. Demand response (DR) programs seek to provide incentives to induce dynamic management of customers' electricity load in response to power supply conditions, for example, reducing their power consumption in response to a peak load warning signal or request from the utility. The National Institute of Standards and Technology (NIST) and the Department of Energy (DoE) have both identified demand response as one of the priority areas for the future smart grid [26], [72]. Further, the National

Assessment of Demand Response Potential report has identified that demand response has the potential to reduce up to 20% of the total peak electricity demand in the U.S. [33].

In this paper, our goal is to highlight that *data centers* represent a particularly promising industry for the adoption of both traditional and advanced demand response programs.

A. Why data centers?

Our focus on data centers stems from the fact that they are particularly well-suited for participation in demand response programs. To see this, note that, first and foremost, data centers represent *very large loads* for the grid. In 2011, they consumed approximately 1.5% of all electricity worldwide. Some individual data centers can consume up to 50 MW, or more [1], [40], [73]. Further, the energy consumption of data centers is *growing quickly*, by approximately 10-12% per year [1], [40], [53]. This growth is crucial for keeping pace with the growth of renewable adoption predicted for the coming years.

Another important aspect about data centers that makes them well-suited for demand response programs is that they are *extremely flexible loads*. Data centers are highly automated and monitored, e.g., the power load and state of IT equipment and cooling facilities can be continuously monitored and panoramically adjusted. For example, recent empirical studies by Lawrence Berkeley National Laboratory (LBNL) has quantified the flexibility in power usage of four data centers under different management approaches [40], [41]. They found that 5% of the load can typically be shed in 5 minutes and 10% of the load can be shed in 15 minutes; and that these can be achieved *without* changes to how the IT workload is handled, i.e., via temperature adjustment and other building management approaches. Further, if workload management approaches are used, the degree of flexibility can be even larger, without additional time needed to shed the load.

A large body of research has recently gone into the design of such approaches to exploit the power consumption flexibility of data centers. Some of the major opportunities for flexible management of data center power demand are the following:

- *Capacity right-sizing*: Over the past decade there has been a large amount of attention given to the design of hardware and algorithms that can adapt energy usage to create "power proportional" systems, that use power in proportion to the utilization of the computing system. Such designs focus on speed-scaling [7], [17], [18], [28], [86], [88], [100], power-capping [15], [34], moving servers into and out of power saving modes [46], [58], [65], [67], [109], and many other features.

- *Load shifting*: Data centers typically have a mixture of workloads. Some are inflexible, i.e., delay intolerant, but many are delay tolerant. For delay-tolerant workload, it is possible to shift work in time to run when renewable or cheaper energy is available, among other things. Many algorithms have been proposed to accomplish this sort of load shifting over time in different circumstances [19], [42], [59], [107].

- *Quality degradation*: In addition to load shifting, another flexibility data centers have is "load shedding", which is typically associated with quality degradation of some form, with possible consideration of quality-of-service (QoS) requirements and service-level-agreements (SLAs). E.g., when serving ads, a data center can use less energy by targeting adds less effectively. This tradeoff can be exploited to reduce energy costs or reduce brown energy usage, among other things [8], [37], [38], [45], [96], [104].

- *Geographical load balancing*: Many Internet-scale systems depend on a number of geographically distributed data centers. Thus, in addition to flexibility within a data center, they have geographical flexibility about the data center location at which they can serve a given workload. This so-called "geographical load balancing" has been shown to be effective in reducing energy costs and improving the efficiency of local renewable energy at data centers, among other things [10], [36], [57], [60], [61], [69], [79], [81], [99].

It is worth noting that many of these approaches can be implemented without impacting the quality of service for IT workloads. However, some certainly can have an impact on quality of service. For instance, moving unused servers into power saving modes or rescheduling delay tolerant workloads may impact response times. Others, e.g., serving ads with reduced effectiveness, or geographical load balancing, may lead to quality degradation. Quantifying such degradation and associated cost, e.g., revenue loss, is an important topic that has received considerable attention, e.g., [16], [22], [60], [61], [77], [89].

In addition to flexibility in the workloads, data centers typically have large scale energy storage on-site in order to provide backup power for their servers [43], [91]. Moreover, they typically also have a backup generator on site in case of extreme failures [63], [73]. Both of these can provide additional opportunities for data centers to have flexibility in the amount of energy that is drawn from the grid.

This diversity of options for achieving flexible energy usage highlighted above, combined with the large peak demands of data centers, makes them an extremely attractive target for demand response programs. To quantify the potential for such participation, we contrast data center demand response and conventional energy storage in following.

B. Data centers demand response versus energy storage

Given the view of demand response as an alternative to conventional energy storage and the attractiveness of data centers as demand response targets, it is natural to try to quantify the potential value of data center demand response participation in terms of the amount of storage it is "equivalent" to. This comparison is the goal of this section. In particular, we ask:

> *How much (optimally placed) energy storage can a data center replace by participating in demand response?*

To provide insight into this question, we focus on the *potential* of data center demand response, and so do not model market factors that lead to inefficiency. Rather, we assume that the load serving entity can call on the data center and storage as needed, and the data center will respond exactly as requested.

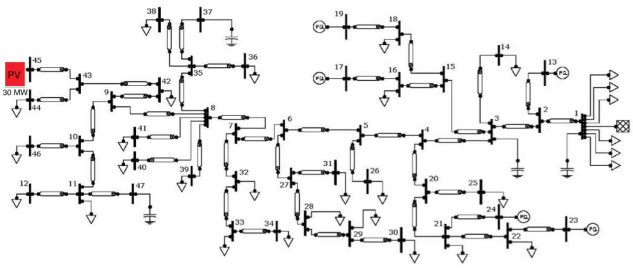


Fig. 1. SCE 47 bus network.

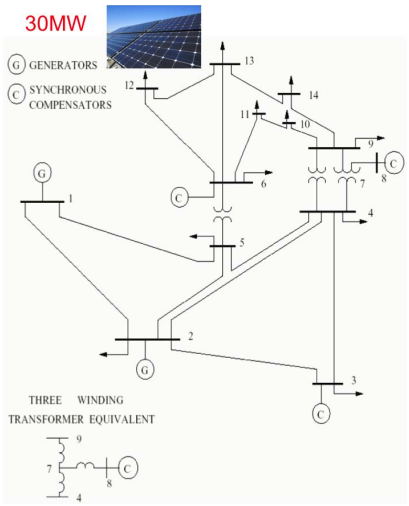


Fig. 2. IEEE 14 bus network.

We present two case studies in this regard. The first follows [62] and the second is novel.

In both cases we study a situation where a distribution network has a large-scale solar installation. Either a large-scale storage facility or a data center helps manage the intermittency of the solar installation. The two case studies differ primarily in the metric of interest for the load serving entity. In the first case study, we focus on limiting the *voltage violation frequency* and in the second we focus on *minimizing cost*.

Throughout, we adopt the models of data center flexibility and energy storage management used in [4], [57], [63], [66], which is based on industry traces from Facebook and Google. Additionally, in order to model a solar installation placed within a power distribution network, we use solar irradiance data from Los Angeles, CA in February 2012 [48] to alter the power load at the bus where the solar (PV) generation is located. Due to space constraints (and for readability), we refer readers to [62] for the details of the experimental setup.

Case study 1:

Our first case study focuses on the potential of data center demand response as a tool for helping a load serving entity reduce voltage violation frequency.

We focus on a distribution network from the Southern California Edison (SCE) utility company. The network includes 47 buses and is pictured in Figure 1. Note that, there is no conventional generation in this distribution network. All power comes from the substation bus, a.k.a., the zero bus, and the solar installation, which is at bus 45 and sized at 30 MW. The demands are taken from SCE load profiles [47], except for the data center, which follows the model in [4], [57], [63], [66].

Given these settings, a significant amount of the solar generation can be transmitted out of the distribution network through the substation bus. However, because we consider a large-scale solar installation, when the installation has near peak generation, the network constraints become binding and voltage violations are common. Note that, the voltage constraint we consider is taken directly from the network tolerance specifications, and is 3%. The number of violations in our simulations are consistent with previous work on these networks, e.g., [31], [32]. The presence of storage or the data center is used to help avoid such violations.

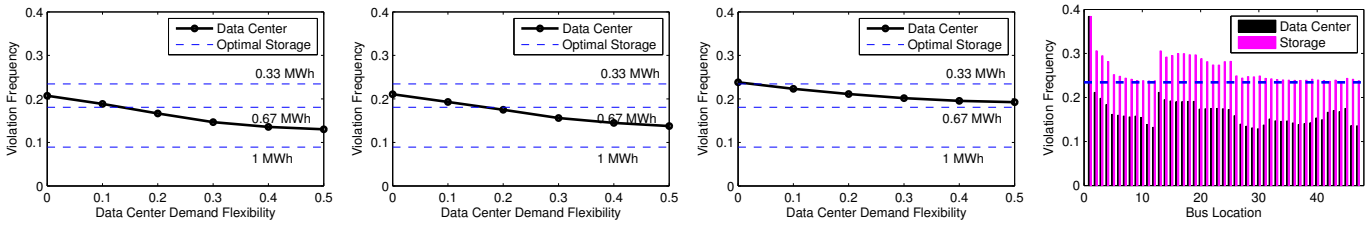
The results that are summarized in Figure 3 highlight that data center demand response has a significant potential. In particular, the comparisons in this plot assume storage with infinite charging speed, i.e., a charging rate of 1, and is thus quite conservative. In Figure 3, we fix the capacity of the data center to 15 MW, which is a representative size for today's IT companies, and then investigate the impact of the degree of data center flexibility, and the placement of the data center. For example, Figures 3(a)-3(c) highlight that the voltage violation rates decrease as data center power demand becomes more flexible. In particular, a 15 MW data center with 20% power demand flexibility placed at the PV location is equivalent to 0.67MWh of optimally-placed storage in the 47 bus distribution network. Further, Figure 3(d) shows that the benefit of data center flexibility is robust to its placement in the power distribution network, i.e., there are very few locations where the effectiveness of the data center drops considerably and many locations that are near-optimal. Figure 3(d) also illustrates that a 15 MW data center is better than 0.33 MWh of conventional energy storage almost uniformly.

Importantly, we can quantify these results monetarily too. Note that, the cost of energy storage is upwards of $500/kWh for lithium-ion batteries (which have small charging rates) and upwards of $5000/kWh for technologies with fast charging rates, such as flywheels. Thus, the flexibility provided by one 30 MW data center is worth upwards of $500,000 - $5,000,000. These numbers are conservative estimates, and grow considerably if a slower charging rate is used in the simulations or if the flexibility of the data center is increased. Thus, *each data center that is not participating in demand response programs represents millions of dollars worth of installed storage capacity that is not being used.*

Case study 2:

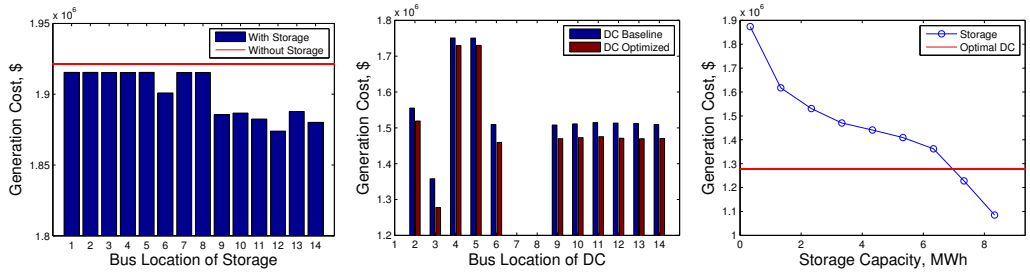
Our second case study focuses on cost rather than voltage violations. In particular, we study the potential of data center demand response as a tool for helping a load serving entity reduce the costs of serving demand in the presence of a large scale renewable installation. We focus on a standard test network: the IEEE 14 bus network. We use a smaller network for this case study due to the computational challenges of the optimal power flow calculation. This is a non-convex optimization that is time-consuming to solve for a single time-step, and we need to solve it across an entire day.

To set up our experiment, we place a 30 MW solar installation at bus 11. Two data center simulations, one with



(a) Data center placed at the optimal storage location (b) Data center placed at the PV location (c) Data center placed at bus 2 (d) Data center vs. storage

Fig. 3. Comparison of a 15 MW data center to large-scale storage in a 47 bus SCE distribution network. (a)-(c) show the violation frequency as a function of the amount of data center flexibility, and compare it to optimally placed storage. (d) shows the violation frequency resulting from a data center with 20% flexibility versus 0.33 MWh of storage, for each location.



(a) Storage placed at each of the 14 buses (b) Data center with baseline usage and data center with optimized usage placed at each of the buses (c) 30MW data center placed at bus 6 compared to a large-scale storage placed at bus 12

Fig. 4. Comparison of a 30 MW data center to large-scale storage in a 14 bus IEEE network. (a) and (b) show the cost benefits of storage and optimized usage data centers. (c) shows the cost resulting from a 30 MW data center with 20% flexibility versus storage with varying capacities.

a constrained baseline usage (average of its usage interval) and another with an optimized usage, are simulated. All other parameters mimic case study 1. Our goal is to strengthen the view that data center demand response can serve the role of energy storage by illustrating the equivalence on the basis of cost (as opposed to voltage violation frequency).

The results are summarized in Figure 4. The key point is that, again, data centers can provide the same service for the load serving entity as a large scale storage installation. Figure 4(a) illustrates the cost benefits of having storage, while figure 4(b) highlights the improvement of an optimized data center versus a baseline constrained one. Figure 4(c) depicts that an optimized 7 MWh large-scale storage is comparable to a 30 MW data center. Thus, *we see that data center demand response is even more valuable here than in case study 1.*

III. Opportunities for Data Center Participation in Demand Response Programs

When illustrating the potential of data center participation in demand response programs in the previous section, we assumed that data centers would adjust their usage (within bounds on flexibility) exactly the way that the grid operator desired. Of course, this is not what happens in practice. However, there are many demand response programs available today that allow the grid operator to extract flexibility from participants through either price signals or direct control signals.

In this section, we survey some of the most promising opportunities for data center participation in electricity market and demand response programs that are available today. We divide the programs into two categories: programs that allow for either "passive" or "active" participation. By passive participation programs, we mean those where participation does not seek to have direct impact on the electricity market, as opposed to active participation programs where participation aims to directly affect the market, e.g., through bidding.

A. Opportunities for passive participation

Passive programs typically use some sort of "smart" pricing approach. That is, consumers are encouraged to *individually* and *voluntarily* manage their loads through the use of pricing signals. These programs come in a variety of forms. The following list shows some of the most common in the U.S.:

- *Time-of-Use Pricing*: Certain times during the day are identified as peak, mid-peak, and off-peak hours, each group having distinct rates for electricity. For example, Portland General Electric Utility has identified 3:00-8:00 PM as peak hours, with peak prices being three times higher than off-peak prices [78].

- *Inclining Block Rates*: Beyond a threshold in the consumer's monthly, daily, or hourly load, the price increases to a higher value [83]. This encourages consumers to keep their load below a certain level at certain times. Inclining block rates are practiced, e.g., by Clatskanie Public Utility for residential users [20] and by Alabama Power for industrial consumers [5].

- *Peak Pricing*: Many utilities also use peak pricing (PP) for large industrial loads, based on their maximum

demand. The maximum demand might be calculated separately for on-peak, off-peak, or mid-peak hours. For example, Riverside Public Utility calculates the maximum demand for each on-peak, off-peak, and mid-peak period based on the maximum average kilowatt input recorded by metering instruments during any 15-minute metered interval in each month [85].

- *Coincident Peak Pricing*: Under coincident peak pricing (CPP), industrial consumers are charged a very high price (often over 200 times higher than the base rate) for usage during the coincident peak hour, i.e., the hour when the most electricity is requested from the utility's wholesale power supplier. These coincident peaks may typically be accompanied by advance but short (e.g., 5 minutes) notice, and are often limited to a maximum number of hours per year. In case of Fort Collins Utilities in Colorado [63], [102], it is common to have about 10 to 12 critical peak warning notices every month.

- *Day-Ahead Pricing*: While time-of-use prices are fixed for several months and limited to only two or three price levels, it is becoming common for many utilities to also offer day-ahead prices (DAPs) that are calculated based on the clearing market prices in the day-ahead market and carry a separate price for each hour of the next day. For example, Ameren Illinois Utilities offer day-ahead prices that are updated daily at 4:30 PM and provide a full table of electricity prices for each hour during the next day [6].

- *Real-Time Pricing*: It some regions, e.g., in Electric Reliability Council of Texas (ERCOT), for consumers to be charged at real-time prices (RTPs) [29]. Such prices are established every 15 minutes based on the clearing market prices in real-time market. Thus, RTPs are not known at the time of usage as they are calculated only after-the-fact. This can cause uncertainties to consumers; however, since RTP charging eliminates the large "insurance premium" that paid for the luxury of purchasing power at flat or pre-determined rates, it can lead to big savings for certain consumers.

B. Opportunities for active participation

In contrast to the opportunities for passive participation, which primarily involve responses to price signals, the market programs we discuss here require active participation in a market via the submission of bids or negotiation. Programs of this type that are appropriate for data centers fall into three categories: wholesale electricity markets, ancillary services markets, and load reduction markets. Each one has multiple participation opportunities, as we explain below.

1) Wholesale markets: While it is typical for consumers to buy electricity from regional retailers, some independent system operators (ISOs), such as ERCOT and California ISO, have recently developed a market that allows consumers to purchase electricity directly from power suppliers by actively participating in one or both of the following markets. These options offer tremendous flexibility to purchase traditional and/or green energy to larger costumers, such as data centers.

- *Bilateral markets*: A medium or large data center can enter a bilateral contract with a power supplier to buy electricity or generation rights under mutual agreements. Bilateral contracts are confidential and flexible. Therefore, data centers can negotiate purchase contracts that can best fit their energy needs given their load characteristics and load control capabilities.

- *Power markets*: A data center may also participate in the wholesale market. A common option for major load entities is to submit "limit order" bids to the day-ahead market. For each hour of the day h, such bids indicate that the data center is willing to buy L_h MW electricity at a price no higher than p_h. Once the day-ahead auction is processed, if the market clearing price at hour h stays below p_h, then the data center purchases the rights to the L_h MW of electricity at hour h and pays the market clearing price. Otherwise, it does not receive the rights to the L_h MW of electricity at hour h and must purchase needed energy in the real-time market at "unknown prices".

2) Ancillary Service markets: Another opportunity that is well-suited to data centers is to participate in ancillary service markets as a "load resource". In fact, many of the existing ancillary service markets, e.g., PJM and ERCOT, allow providing a portion (e.g., 20%, in case of PJM) of their ancillary services from load resources. Ancillary services are defined as the services necessary to support the transmission of energy to loads while maintaining reliable operation and security of the electricity transmission system.

Balancing supply and demand can be achieved by either adjusting generation or adjusting consumption. Therefore, payments for load reductions to load resources are equal, dollar for dollar, to that which suppliers are paid for increasing generation. In fact, similar to generators, the "value" of a load resource (e.g., a large data center) depends on three factors: (i) how quickly it can respond to change (reduce or increase) its load; (ii) the cost at which a load resource is willing to adjust its load; and (iii) the market condition at which the service was offered. Accordingly, there are different ancillary services that could be offered by load resources based on their capabilities. In PJM and ERCOT, such services differ in response time and are as follows [29], [75]:

- *Spinning reserves:* In this service, a command to interrupt or reduce the load comes either from an on-site under-frequency relay (UFR) or through a (*10 minutes-ahead or shorter*) notice signal from the ISO. The load resource is then required to provide holding service for at least 15 minutes and up to multiple hours. The spinning reserve service is also referred to as "responsive reserve service".

- *Non-spinning reserves:* Non-spinning reserves provide the same service as spinning reserves, but are not required to respond to notices as quickly, i.e., signals arrive with *30-minutes notice* typically.

- *Regulation services:* When offering regulation service, a flexible load (such as data center) needs to respond to up/down signals that arrive, e.g., *every 4 or 10 seconds*, by decreasing/increasing the load accord-

ingly, while meeting rigorous performance monitoring criteria. Regulation can be done at different resolutions. For example, in PJM, there are two, Reg A (traditional) and Reg D (dynamic), regulation signals [75]. Reg D command signals fluctuate more severely. Accordingly, there is a higher payment for offering dynamic regulation.

In general, making decisions to offer ancillary service is very difficult. However, if it is done properly, it has the potential to bring major financial benefits to data centers, in addition to helping the grid. To be qualified as a load resource, a data center must (i) meet a minimum flexible load capacity (e.g., 1 MW in ERCOT), (ii) install real-time telemetry systems, and (iii) pass and maintain high scores in "performance tests".

The payments for participation in such programs are quite complicated. We give a brief overview in the following.

- *Load resource payments*: Load resources that offer ancillary services typically receive two types of payments. The first payment is the *"capacity payment"*, which is made simply for being available. The second payment is the *"operation payment"*, which is made only if the service was actually called. For the responsive and non-spinning reserves, this payment is typically calculated based on the *locational marginal price* (LMP) at the power grid bus where the load resource is located. For regulation services, additional payments are made based on *"mileage"* for each regulation signal type [75], which is combined with several factors such as LMP, "benefit factor" (that indicates the scarcity of load and generation resources to perform regulation), "historical performance score" of the load resource, and the total regulation capacity that is offered by the load resource.

- *Performance evaluation*: While assessing the performance of reserve services is typically simple, the performance evaluation for regulation services requires advanced monitoring and analysis. For example, PJM evaluates regulation performance based on scores on *"delay"*, *"correlation"*, and *"precision"* [76]. The Delay Score quantifies the delay between the regulation signal and changes in demand. The Correlation Score measures the accuracy in matching the regulation signal, using the correlation between regulation and response signals. The Precision Score is calculated as an hourly average of the difference between the regulation and response signals over 10 seconds sampling intervals. The final performance score is calculated as a weighted summation of all three scores. Maintaining a minimum (e.g., 75%) score is needed to stay qualified to offer regulation services.

- *Bidding process*: The bids for offering ancillary services are submitted to ancillary service markets. Various information must be included in the bid. For example, for regulation services, the capacity and the regulation type (traditional or dynamic) should be indicated. The financial element of the bid could be *"cost-based"* or *"price-based"*. The former parameterizes the service cost function, e.g., in terms of start-up and incremental costs for local generators. The latter is in the form of price schedules that indicate the price of offering the service at each time of operation.

3) Voluntary Load Reduction: A third option well-suited for data centers is to offer some voluntary services to regional grid operators. For example, in ERCOT, industrial consumers can offer "voluntary load reduction" services to regional operators, called Qualified Scheduling Entities (QSEs). There are at least two key distinctions between offering load reduction to QSEs and offering ancillary services to ISOs that lead to important differences for data center management. First, such services are voluntary and usually guarantee only best-effort services, thus participation carries little or no risk. In turn, they typically have lower payments. Second, they do not require bidding and have flexible contracts. Thus, a potential load resource such as data center will need to negotiate with its corresponding QSE to settle down the terms of the contract.

IV. Challenges that limit data center participation in demand response

The previous sections have highlighted the potential for data center demand response and the opportunities data centers have for participation. It is important to emphasize that data center participation in demand response programs truly has the potential to be a "win-win": data centers provide a significant service to grid operators *and* demand response programs provide a significant revenue source for data centers.

However, despite this potential "win-win" opportunity, data centers today are largely non-participants in the demand response programs we discuss above. The reasons for this are not mysterious. There are a number of significant challenges that lead to this unfortunate fact. Below, we outline some of these biggest reasons. Then, in the next section, we discuss recent research progress in the academic and industrial research communities that is beginning to alleviate these challenges.

Challenge 1: Regulation and market maturity

First and foremost, it is important to emphasize that, though we have outlined a large number of participation opportunities for data centers in demand response programs, many of these programs are not available to data centers in markets today. While some utilities have been quick to move to adjust regulations to allow greater participation in market programs, many have been quite slow. As a result, in any given area, the opportunities for data center demand response participation may be limited to simple, traditional smart pricing programs such as coincident peak pricing which, as we discuss next, are not well-suited for the risk tolerance of data centers.

Challenge 2: Risk management

Data centers are typically in the business of maximizing uptime and performance, and energy issues are certainly secondary to maintaining strong guarantees about these primary measures. However, participation in demand response programs always comes with some risk. This risk may be purely financial, e.g., in passive participation programs, or it may have the possibility of uptime/performance degradations, e.g., in active participation programs. As a result, risk management is a crucial issue for data center participation in demand response programs. Taking a huge financial/performance hit because the grid sends a price/control signal at the same point when the

data center is heavily loaded is a serious concern that limits data center participation in current market programs. In fact, for exactly this reason data centers prefer to negotiate long term energy contracts with fixed usage prices.

Challenge 3: Who has control?

An active debate within the demand response field is that of who should have control? Grid operators would like to have a guaranteed response when they ask for it; which leads to "direct load control" programs for which the grid sends a signal to a controller of the program participant. However, of course, this is not always acceptable to participants. In particular, such programs are inappropriate for data centers given the risk management issues discussed above. The other extreme alternative is "prices-to-devices" where real-time prices are conveyed to participants; however such programs typically require huge price variation in order to extract desired responses. Again, this volatility is not acceptable given the risk tolerance of data centers. Thus, other programs must be developed in order to facilitate data center participation.

Challenge 4: Market complexity

Financially, the active participation programs we have described have a huge potential for data centers. However, as we have discussed, participation in these programs is highly regulated and the bidding necessary to extract profits is something that is typically difficult to automate and incorporate into a data center management system. This complexity has, to this point, prevented data centers from entering these markets despite the financial opportunities.

Challenge 5: Market power

The challenges that we have outlined so far relate to data center participation. However, there are also significant challenges on the grid operator side. One that is particularly salient is the potential for data centers to manipulate market prices. In particular, as we have discussed, data centers are very large loads. They can make up 20-50% of the load on their distribution circuit. In such situations, if they participate aggressively in some of these market programs there is a significant potential for them to wield market power to manipulate prices in their favor. Given that many of these markets have been designed for situations in which many small loads all act as price-takers, grid operators are rightfully nervous about loosening regulations to allow data center participation.

V. Recent progress in data center demand response

Given the challenges that remain before data center demand response participation can realize its potential, there are clearly many important research questions to address. To that end, a new field is emerging at the intersection of data center management and power systems that focuses on facilitating the interaction of data centers in demand response programs. In the following, we survey some of the progress that has been made toward addressing the challenges we outlined in the previous section. Note that, though progress has been made, it is clear that many, significant challenges are yet to be addressed.

We organize the progress made to this point into two categories: (A) progress toward the improved management of data centers to facilitate participation in demand response programs; and (B) progress toward the design of new market programs that are appropriate for data center participation.

A. Managing data center participation in demand response

The task of managing data center participation in a demand response program is clearly a difficult one; however, because of the large literature on energy-efficient data centers that has emerged over the past decade, there are many tools that have already been well-developed at this point. In particular, techniques for right-sizing, load shifting, quality degradation, etc., are developed and, sometimes, used in practice already. However, the challenge of how to use them to optimize participation in demand response programs is still unsolved.

In particular, different demand response markets require very different strategies. Classically, much of the academic work on energy-efficient data centers has focused on time-of-use pricing, and so there are many strategies available for such programs, e.g., [12], [19], [34], [46], [49], [51], [55], [58], [59], [66], [67], [94], [107], [109]. The algorithmic challenges in such designs often stem from the unpredictability of workload and the costs associated with switching the state of servers.

More generally though, there are many other options for demand response programs which can provide significantly larger financial incentives for data centers. For example, it is often beneficial for data centers to hedge long-term energy contracts with participation in spot-markets, thus creating a challenging online, multi-time scale optimization problem. Designs have started to emerge for optimizing such contracts [23], [24], [71], [80], [82], [108].

Another popular option for demand response is coincident peak pricing programs. Such programs provide a challenge for data center management since there is significant uncertainty about when coincident peak warnings will be sent to the data center, thus signaling a reduction. Recent work has looked at using online, robust optimization as a tool for managing participation in such programs [11], [12], [52], [63], [87], [93].

The programs we have discussed so far are all passive. Participation in active demand response programs is much more challenging, and has only recently begun to be studied. For example, recent papers have looked at managing data center participation in regulation markets and ancillary service markets [2], [3], [13]–[15], [35].

In addition to the details of the particular program, there are key challenges that demand response can create within data centers. For example, many data centers are multi-tenant, i.e., they rent space to many different tenants. In such situations, the data center operator does not have control over the computing resources and so when a demand response signal is received, it cannot manage the response directly and must find a way to encourage the tenants to respond appropriately. Some recent work has looked at designing mechanisms for this setting [84].

Another level of complexity on top of all the issues we have discussed so far is the fact that data centers often have local resources such as energy storage, renewable energy, and/or backup generators on-site. Each of these adds additional uncertainty and complexity to the participation decisions discussed above, and each has been studied by recent work [36], [37], [63], [64], [90], [92].

B. Design of market programs appropriate for data centers

While significant progress has been made on developing tools and algorithms for facilitate data centers participation in demand response programs, it is clear that, in the long term, the development of new market programs are crucial to efficiently extract data centers flexibility. However, it is not at all clear yet what form these new market programs should take.

There are multiple tradeoffs at play in the design of new market programs. Should the new programs be passive or active? How much control should the load serving entity wield versus the data center? What time-scale should data centers be encouraged to provide flexibility over? These and many other questions are at the heart of the emerging research on market designs for data center demand response. For example, the information communicated between utility company and data centers, e.g., real-time prices and regulating signals, can be viewed as the approach to decouple the global optimization problem into two subproblems: one for utility company and one for data centers. Optimal design of such mechanism is a highly important topics [39].

One key issue that has emerged as crucial in the design of new market programs is the market power that data centers wield. As we have already highlighted, data centers can make up a significant proportion of the load on a given distribution circuit, and thus they have the potential to significantly manipulate prices if care is not taken in design.

This worry is particularly salient given the typical "price-taker" assumption that goes into the design of most demand response programs. Clearly data centers need not be price-takers. However, quantifying the potential for market power is a difficult task, and only recently have market power metrics that incorporate transmission constraints begun to emerge [9], [54], [101], [105]. Noticeably, none of these metrics are designed for assessing market power on distribution networks.

Thus, there seems to be a tradeoff between pricing approaches and bid-based approaches in terms of market power versus prediction error. Specifically, bid-based approaches generally suffer if a participant has market power, e.g., [50], [97], [106], while pricing-based approaches require predicting the flexibility of participants in order to set prices efficiently, e.g., [21], [44], [56], [70], [98]. To this point, it is not yet clear which is more appropriate for data center demand response programs.

Finally, the above discussion has focused entirely on a single data center. To this point, there are no existing demand response programs that are designed to extract geographic flexibility. Such programs could be of crucial importance in areas where large-scale solar installations stress circuits across different regions in a load serving entity [68], [95].

VI. Concluding remarks

In this survey we have laid out the potential benefits that come from integration of data centers into demand response programs. We have also highlighted the challenges that are currently preventing most data centers from active participation in these programs, and the ongoing research efforts in academia and industry focused on overcoming these challenges.

We would like to conclude by highlighting that it is our perspective that none of these challenges are overwhelming, and that when grid operators look at data centers they should view them as large scale energy storage installations that are sitting unused due to a lack of appropriate market programs. In particular, *each data center represents millions of dollars worth of unused fast-response storage-equivalent capacity*.

Given the lack of cost-effective large-scale energy storage, we believe that the development of market programs to extract flexibility from data centers is crucial for easing the incorporation of renewable energy into the grid as well as conducting peak-load shaving. We hope that this survey serves to highlight the research challenges that remain before the potential of large scale data center demand response can be realized.

Acknowledgment

This work was supported by NSF grants CCF 0830511, CNS 0911041, CNS 1319798, ECCS 1253516, ECCS 1307756 and CNS 0846025, DoE grant DE-EE0002890, ARO MURI grant W911NF-08-1-0233, Microsoft Research, Bell Labs, the Lee Center for Advanced Networking, and ARC grant FT0991594. The authors would like to thank Ayse Coskun and Bhuvan Urgaonkar for their valuable discussion.

References

[1] EPA report to congress on server and data center energy efficiency. 2007.

[2] D. Aikema, R. Simmonds, and H. Zareipour. Data centres in the ancillary services market. In *Green Computing Conference (IGCC), 2012 International*, pages 1–10. IEEE, 2012.

[3] B. Aksanli and T. Rosing. Providing regulation services and managing data center peak power budgets. In *Proc. of the IEEE Design, Automation and Test in Europe Conference and Exhibition*, Dresden, Germany, Mar. 2014.

[4] B. Aksanli, J. Venkatesh, L. Zhang, and T. Rosing. Utilizing green energy prediction to schedule mixed batch and service jobs in data centers. *ACM SIGOPS Operating Systems Review*, 45(3):53–57, 2012.

[5] Alabama Power - A Southern Company. Incemental Load Pricing. http://www.alabamapower.com/business/pricing-rates/pdf/ILD.pdf, Apr. 2011.

[6] Ameren Energy. Day-ahead and Real Time Electricity Prices. https://www2.ameren.com/RetailEnergy/realtimeprices.aspx, Dec. 2012.

[7] L. L. H. Andrew, M. Lin, and A. Wierman. Optimality, fairness and robustness in speed scaling designs. In *Proc. of ACM Sigmetrics*, 2010.

[8] W. Baek and T. M. Chilimbi. Green: a framework for supporting energy-conscious programming using controlled approximation. In *ACM Sigplan Notices*, volume 45, pages 198–209. ACM, 2010.

[9] D. W. Cai and A. Wierman. Inefficiency in forward markets with supply friction. In *Proc. of IEEE CDC*, 2013.

[10] J. Camacho, Y. Zhang, M. Chen, and D. Chiu. Balance your bids before your bits: The economics of geographic load-balancing. In *Proc. of ACM e-Energy*, 2014.

[11] P. Cappers, C. Goldman, and D. Kathan. Demand response in us electricity markets: Empirical evidence. *Energy*, 35(4):1526–1535, 2010.

[12] J. Chase. Demand response for computing centers. In *"The Green Computing Book: Tackling Energy Efficiency at Large Scale"*. CRC Press, 2014.

[13] H. Chen, M. Caramanis, and A. K. Coskun. The data center as a grid load stabilizer. *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014.

[14] H. Chen, A. K. Coskun, and M. C. Caramanis. Real-time power control of data centers for providing regulation service. *Proc. 52nd CDC*, 2013.

[15] H. Chen, C. Hankendi, M. C. Caramanis, and A. K. Coskun. Dynamic server power capping for enabling data center participation in power markets. In *Proceedings of the International Conference on Computer-Aided Design*, pages 122–129. IEEE Press, 2013.

[16] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. Quantifying skype user satisfaction. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 399–410. ACM, 2006.

[17] L. Chen, N. Li, and S. H. Low. On the interaction between load balancing and speed scaling. In *ITA Workshop*, 2011.

[18] W. Chen, D. Huang, A. A. Kulkarni, J. Unnikrishnan, Q. Zhu, P. Mehta, S. Meyn, and A. Wierman. Approximate dynamic programming using fluid and diffusion approximations with applications to power management. In *Proceedings of CDC*, pages 3575–3580, 2009.

[19] Y. Chen, D. Gmach, C. Hyser, Z. Wang, C. Bash, C. Hoover, and S. Singhal. Integrated management of application performance, power and cooling in data centers. In *Proc. of NOMS*, 2010.

[20] Clatskanie People Utility District. Rate Schedules Summary. http://www.clatskaniepud.com/Rate Schedules.htm, June 2012.

[21] A. J. Conejo, J. M. Morales, and L. Baringo. Real-time demand response model. *IEEE Transactions on Smart Grid*, 1(3):236–242, 2010.

[22] A. Croll and S. Power. How web speed affects online business kpis, 2009.

[23] W. Deng, F. Liu, H. Jin, and C. Wu. Smartdpss: cost-minimizing multi-source power supply for datacenters with arbitrary demand. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*, pages 420–429. IEEE, 2013.

[24] W. Deng, F. Liu, H. Jin, C. Wu, and X. Liu. Multigreen: cost-minimizing multi-source datacenter power supply with online control. In *Proceedings of the fourth international conference on Future energy systems*, pages 149–160. ACM, 2013.

[25] Department of Energy. 20% wind energy by 2030. http://www.20percentwind.org/.

[26] Department of Energy. The smart grid: An introduction.

[27] Department of Energy. Installed wind capacity. http://www.windpoweringamerica.gov/wind_installed _capacity.asp, June 2012.

[28] M. Elahi, C. Williamson, and P. Woelfel. Decoupled speed scaling: Analysis and evaluation. *Performance Evaluation*, 2013.

[29] Electric Reliability Council of Texas. Load Participation in the ERCOT Nodal Market. Prepared by the Demand-Side Working Group of the ERCOT Wholesale Market Subcommittee, Version N 1.0, June 2007.

[30] X. Fang, S. Misra, G. Xue, and D. Yang. Smart gridthe new and improved power grid: a survey. *Communications Surveys & Tutorials, IEEE*, 14(4):944–980, 2012.

[31] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy. Inverter var control for distribution systems with renewables. In *IEEE SmartGridComm*, pages 457–462, 2011.

[32] M. Farivar, R. Neal, C. Clarke, and S. Low. Optimal inverter var control in distribution systems with high pv penetration. In *IEEE Power and Energy Society General Meeting*, pages 1–7, 2012.

[33] Federal Energy Regulatory Commission. National assessment of demand response potential. 2009.

[34] A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah. Minimizing data center SLA violations and power consumption via hybrid resource provisioning. In *Proc. of IGCC*, 2011.

[35] M. Ghamkhari and H. Mohsenian-Rad. Data centers to offer ancillary services. In *Proc. of the IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2012.

[36] M. Ghamkhari and H. Mohsenian-Rad. Optimal integration of renewable energy resources in data centers with behind-the-meter renewable generator. In *Proc. of IEEE International Conference on Communications (ICC)*, 2012.

[37] M. Ghamkhari and H. Mohsenian-Rad. Energy and performance management of green data centers: a profit maximization approach. *IEEE Trans. on Smart Grid*, 4(2):1017–1025, 2013.

[38] M. Ghamkhari and H. Mohsenian-Rad. Profit maximization and power management of green data centers supporting multiple slas. In *Proc. of IEEE International Conference on Computing, Networking and Communications (ICNC)*, 2013.

[39] G. Ghatikar and R. Bienert. Smart grid standards and systems interoperability: a precedent with openadr. In *Proceedings of the Grid Interop Forum*, 2011.

[40] G. Ghatikar, V. Ganti, N. Matson, and M. Piette. Demand response opportunities and enabling technologies for data centers: Findings from field studies. 2012.

[41] G. Ghatikar, M. A. Piette, S. Fujita, A. McKane, J. Han, A. Radspieler, K. Mares, and D. Shroyer. Demand response and open automated demand response opportunities for data centers. *California Energy Commission, PIER Program and Pacific Gas and Electric Company (PG&E)*, 2010.

[42] D. Gmach, J. Rolia, C. Bash, Y. Chen, T. Christian, A. Shah, R. Sharma, and Z. Wang. Capacity planning and power management to exploit sustainable energy. In *Proc. of CNSM*, 2010.

[43] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar. Aggressive datacenter power provisioning with batteries. *ACM Transactions on Computer Systems (TOCS)*, 31(1):2, 2013.

[44] H. Mohsenian-Rad, V. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia. Autonomous Demand Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid. *IEEE Transactions on Smart Grid*, 1(3):320–331, Dec. 2010.

[45] Y. He, S. Elnikety, J. Larus, and C. Yan. Zeta: scheduling interactive services with partial execution. In *Proceedings of the Third ACM Symposium on Cloud Computing*, page 12. ACM, 2012.

[46] J. Heo, P. Jayachandran, I. Shin, D. Wang, T. Abdelzaher, and X. Liu. Optituner: On performance composition and server farm energy minimization application. *Parallel and Distributed Systems, IEEE Transactions on*, 22(11):1871–1878, 2011.

[47] https://www.sce.com/wps/portal/home/regulatory/load-profiles.

[48] http://www.nrel.gov/midc/lmu/.

[49] D. Irwin, N. Sharma, and P. Shenoy. Towards continuous policy-driven demand response in data centers. In *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, pages 19–24. ACM, 2011.

[50] R. Johari and J. N. Tsitsiklis. Parameterized supply function bidding: Equilibrium and efficiency. *Operations research*, 59(5):1079–1089, 2011.

[51] C. Kelly, A. Ruzzelli, and E. Mangina. Using electricity market analytics to reduce cost and environmental impact. In *Green Technologies Conference, 2013 IEEE*, pages 414–421. IEEE, 2013.

[52] S. Kiliccote, M. A. Piette, and D. Hansen. Advanced controls and communications for demand response and energy efficiency in commercial buildings. 2006.

[53] J. Koomey. Growth in data center electricity use 2005 to 2010. *Oakland, CA: Analytics Press. August*, 1:2010, 2011.

[54] Y. Y. Lee, R. Baldick, and J. Hur. Firm-based measurements of market power in transmission-constrained electricity markets. *IEEE Transactions on Power Systems*, 26(4):1962–1970, Nov. 2011.

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

[56] N. Li, L. Chen, and S. H. Low. Optimal demand response based on utility maximization in power networks. In *IEEE Power and Energy Society General Meeting*, pages 1–8, 2011.

[57] M. Lin, Z. Liu, A. Wierman, and L. Andrew. Online algorithms for geographical load balancing. In *Proc. of IGCC*, 2012.

[58] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *Proc. of INFOCOM*, 2011.

[59] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser. Renewable and cooling aware workload management for sustainable data centers. In *Proc. of ACM Sigmetrics*, 2012.

[60] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew. Geographical load balancing with renewables. In *Proc. ACM GreenMetrics*, 2011.

[61] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew. Greening geographical load balancing. In *Proc. ACM Sigmetrics*, 2011.

[62] Z. Liu, I. Liu, S. Low, and A. Wierman. Pricing data center demand response. In *Proc. ACM Sigmetrics*, 2014.

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

[64] L. Lu, J. Tu, C.-K. Chau, M. Chen, and X. Lin. Online energy generation scheduling for microgrids with intermittent energy sources and co-generation. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, pages 53–66. ACM, 2013.

[65] T. Lu, M. Chen, and L. L. Andrew. Simple and effective dynamic provisioning for power-proportional data centers. *Parallel and Distributed Systems, IEEE Transactions on*, 24(6):1161–1171, 2013.

[66] A. H. Mahmud and S. Ren. Online capacity provisioning for carbon-neutral data center with demand-responsive electricity prices. *ACM SIGMETRICS Performance Evaluation Review*, 41(2):26–37, 2013.

[67] D. Meisner, C. Sadler, L. Barroso, W. Weber, and T. Wenisch. Power management of online data-intensive services. In *Proc. of ISCA*, 2011.

[68] A.-H. Mohsenian-Rad and A. Leon-Garcia. Coordination of cloud computing and smart power grids. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 368–372. IEEE, 2010.

[69] A.-H. Mohsenian-Rad and A. Leon-Garcia. Energy-information transmission tradeoff in green cloud computing. In *Proc. of IEEE Conference on Global Communications (GLOBECOM10)*, Miami, FL, Dec. 2010.

[70] A.-H. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2):120–133, 2010.

[71] J. Nair, S. Adlakha, and A. Wierman. Energy procurement strategies in the presence of intermittent sources. In *Proceedings of ACM Sigmetrics*, 2014.

[72] National Institute of Standards and Technology. NIST framework and roadmap for smart grid interoperability standards. NIST Special Publication 1108, 2010.

[73] NY Times. Power, Pollution and the Internet.

[74] D. of Energy. The smart grid: An introduction. *http://energy.gov/oe/downloads/smart-grid-introduction-0*, 2008.

[75] Pennsylvania Jersey Maryland Interconnect. PJM Manual 11: Energy and Ancillary Services Market Operations, Oct. 2012.

[76] Pennsylvania Jersey Maryland Interconnect. PJM Manual 12: Balancing Operations, July 2012.

[77] E. T. Peterson. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram, 2004.

[78] Portland General Electric. Time of Use Pricing. http://www.portlandgeneral.com, Dec. 2012.

[79] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. In *Proc. of ACM Sigcomm*, 2009.

[80] R. Rajagopal, E. Bitar, P. Varaiya, and F. Wu. Risk-limiting dispatch for integrating renewable power. *International Journal of Electrical Power & Energy Systems*, 44(1):615–628, 2013.

[81] L. Rao, X. Liu, L. Xie, and W. Liu. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *Proc. of INFOCOM*, 2010.

[82] L. Rao, X. Liu, L. Xie, and Z. Pang. Hedging against uncertainty: A tale of internet data center operations under smart grid environment. *Smart Grid, IEEE Transactions on*, 2(3):555–563, 2011.

[83] P. Reiss and M. White. Household electricity demand, revisited. *Review of Economic Studies*, 72(3):853–883, July 2005.

[84] S. Ren and M. A. Islam. A first look at colocation demand response. In *Proc. of ACM GreenMetrics*, 2014.

[85] Riverside Public Utility. Large General and Industrial Service. City of Riverside - Public Utilities Department, Council Resolution No. 22277, 2011.

[86] K. Son and B. Krishnamachari. Speedbalance: speed-scaling-aware optimal load balancing for green cellular networks. In *INFOCOM, 2012 Proceedings IEEE*, pages 2816–2820. IEEE, 2012.

[87] K. Spees and L. B. Lave. Demand response and electricity market efficiency. *The Electricity Journal*, 20(3):69–85, 2007.

[88] R. Stanojevic and R. Shorten. Distributed dynamic speed scaling. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–5. IEEE, 2010.

[89] T. M. Tripp, Y. Grégoire, and S. Business. When unhappy customers strike back on the internet. 2011.

[90] J. Tu, L. Lu, M. Chen, and R. Sitaraman. Dynamic provisioning in next-generation data centers with on-site power production. 2013.

[91] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 221–232. ACM, 2011.

[92] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *Proc. ACM Sigmetrics*, 2011.

[93] C. Wang, B. Urgaonkar, Q. Wang, and G. Kesidis. A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing.

[94] C. Wang, B. Urgaonkar, Q. Wang, G. Kesidis, and A. Sivasubramaniam. Data center power cost optimization via workload modulation. In *Proceedings of the 6th IEEE/ACM International Conference on Uitlity and Cloud Computing (UCC'13) (short paper)*, 2013.

[95] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad. Exploring smart grid and data center interactions for electric power load balancing. *ACM SIGMETRICS Performance Evaluation Review*, 41(3):89–94, 2014.

[96] K. Wang, M. Lin, F. Ciucua, A. Wierman, and C. Lin. Characterizing the impact of the workload on the value of dynamic resizing in data centers. In *INFOCOM, 2013 Proceedings IEEE*, pages 515–519. IEEE, 2013.

[97] P. Wang, L. Rao, X. Liu, and Y. Qi. D-pro: dynamic data center operations with demand-responsive electricity prices in smart grid. *Smart Grid, IEEE Transactions on*, 3(4):1743–1754, 2012.

[98] R. Wang, N. Kandasamy, and C. Nwankpa. Data centers as demand response resources in the electricity market: Some preliminary results. In *Intl. Workshop on Feedback Computing*, 2012.

[99] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford. Donar: decentralized server selection for cloud services. In *Proc. of ACM Sigcomm*, 2010.

[100] A. Wierman, L. L. H. Andrew, and A. Tang. Power-aware speed scaling in processor sharing systems. In *Proc. of INFOCOM*, 2009.

[101] C. Wu, S. Bose, A. Wierman, and H. Mohsenian-Rad. A unifying approach to assess market power in deregulated electricity markets. *Under preparation*, 2014.

[102] www.fcgov.com/utilities/business/rates/electric/coincident-peak.

[103] H. Xu and B. Li. Cost efficient datacenter selection for cloud services. In *Proc. of ICCC*, 2012.

[104] H. Xu and B. Li. Reducing electricity demand charge for data centers with partial execution. In *Proceedings of ACM e-Energy*, 2014.

[105] L. Xu and R. Baldick. Transmission-constrained residual demand derivative in electricity markets. *IEEE Transactions on Power Systems*, 22(4):1563–1573, Nov. 2007.

[106] Y. Xu, N. Li, and S. Low. Demand response with parameterized supply function bidding.

[107] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely. Data centers power reduction: A two time scale approach for delay tolerant workloads. In *Proc. of INFOCOM*, pages 1431–1439, 2012.

[108] L. Yu, T. Jiang, Y. Cao, and Q. Zhang. Risk-constrained operation for internet data centers in deregulated electricity markets. *Parallel and Distributed Systems, IEEE Transactions on*, 25(5):1306–1316, 2014.

[109] Q. Zhang, M. Zhani, Q. Zhu, S. Zhang, R. Boutaba, and J. Hellerstein. Dynamic energy-aware capacity provisioning for cloud computing environments. In *ICAC*, 2012.